Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **501,** 4184–4213 (2021) doi:10.1093/mnras/staa3691
Advance Access publication 2020 November 28

Testing physical models for cosmic ray transport coefficients on galactic scales: self-confinement and extrinsic turbulence at ∼GeV energies

Philip F. Hopkins [1]★ Jonathan Squire [2] T. K. Chan [3,4] Eliot Quataert [5] Suoqing Ji [1] Dušan Kereš[2] and Claude-André Faucher-Giguère [6]

[1]*TAPIR, California Institute of Technology, Mailcode 350-17, Pasadena, CA 91125, USA*
[2]*Physics Department, University of Otago, 730 Cumberland St., Dunedin 9016, New Zealand*
[3]*Department of Physics, Center for Astrophysics and Space Science, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*
[4]*Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK*
[5]*Theoretical Astrophysics Center, Department of Astronomy, University of California Berkeley, Berkeley, CA 94720, USA*
[6]*Department of Physics and Astronomy and CIERA, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA*

Accepted 2020 November 5. Received 2020 October 5; in original form 2020 February 5

ABSTRACT

The microphysics of ∼GeV cosmic ray (CR) transport on galactic scales remain deeply uncertain, with almost all studies adopting simple prescriptions (e.g. constant diffusivity). We explore different physically motivated, anisotropic, dynamical CR transport scalings in high-resolution cosmological Feedback In Realistic Environment (FIRE) simulations of dwarf and $\sim L_*$ galaxies where scattering rates vary with local plasma properties motivated by extrinsic turbulence (ET) or self-confinement (SC) scenarios, with varying assumptions about e.g. turbulent power spectra on un-resolved scales, Alfvén-wave damping, etc. We self-consistently predict observables including γ-rays (L_γ), grammage, residence times, and CR energy densities to constrain the models. We demonstrate many non-linear dynamical effects (not captured in simpler models) tend to enhance confinement. For example, in multiphase media, even allowing arbitrary fast transport in neutral gas does not substantially reduce CR residence times (or L_γ), as transport is rate-limited by the ionized WIM and 'inner CGM' gaseous halo (10^4–10^6 K gas within $\lesssim$ 10–30 kpc), and L_γ can be dominated by trapping in small 'patches'. Most physical ET models contribute negligible scattering of ∼1–10 GeV CRs, but it is crucial to account for anisotropy and damping (especially of fast modes) or else scattering rates would violate observations. We show that the most widely assumed scalings for SC models produce excessive confinement by factors $\gtrsim$100 in the warm ionized medium (WIM) and inner CGM, where turbulent and Landau damping dominate. This suggests either a breakdown of quasi-linear theory used to derive the CR transport parameters in SC, or that other novel damping mechanisms dominate in intermediate-density ionized gas.

Key words: instabilities – plasmas – cosmic rays – ISM: structure – galaxies: evolution – gamma-rays: galaxies.

1 INTRODUCTION

Understanding the propagation or bulk transport of cosmic rays (CRs) through the interstellar, circumgalactic, and intergalactic medium (ISM, CGM, IGM) remains a fundamental and unsolved problem of critical importance for high-energy particle physics, plasma physics, and the astrophysics of star and galaxy formation. In the Milky Way (MW), and (probably) most dwarf and star-forming galaxies, the CR energy density and pressure are dominated by relatively low-energy ∼GeV protons, which are likely accelerated in supernovae (SNe) remnants (with ∼ 10 per cent of the ejecta kinetic energy going into CRs; Bell 2004). These ∼GeV CRs are therefore the most important population governing the interaction of CRs with gas dynamics, heating and cooling of the ISM, gamma-ray emissivities of galaxies, star and galaxy formation, and the excitation of various 'streaming instabilities' and resonant Alfvén waves in the plasma (Kulsrud & Pearce 1969; Mannheim & Schlickeiser 1994; Enßlin et al. 2007; Guo & Oh 2008). There has been a tremendous amount of both analytic (Everett et al. 2008; Socrates, Davis & Ramirez-Ruiz 2008; Dorfi & Breitschwerdt 2012; Mao & Ostriker 2018) and numerical (Jubelgas et al. 2008; Uhlig et al. 2012; Wiener, Zweibel & Oh 2013b; Salem & Bryan 2014; Pakmor et al. 2016; Simpson et al. 2016; Ruszkowski, Yang & Zweibel 2017; Girichidis et al. 2018) work studying these effects. Recent work on galactic scales has argued ∼GeV CRs can play an important role, in particular, in the CGM, by suppressing accretion on to low-redshift $\sim L_*$ galaxies, launching or re-accelerating galactic outflows in these systems, and strongly modifying the phase structure of cool and warm absorption systems (Salem, Bryan & Corlies 2016; Butsky & Quinn 2018; Chan et al. 2019; Hopkins et al. 2020b; Ji et al. 2020; Su et al. 2020).

The transport of these low-energy CRs is *especially* uncertain because (1) there are limited direct observational constraints; (2) the gyro-radii of such CRs are extremely small ($\lesssim$ 1 au), much smaller than observationally resolved scales in most of the MW ISM (let alone other galaxies); (3) the 'back-reaction' of the magnetic fields and gas from CRs (e.g. excitation of Alfvén waves via gyro-resonant instabilities) is maximized around this energy scale because this is

★ E-mail: phopkins@caltech.edu

where the CR energy density is maximized, and can strongly non-linearly alter the propagation of the CRs, i.e. they are 'self-confined'; and (4) the structure of the ISM/CGM in which the CRs propagate is uncertain.

For example, in most of the previous literature, constraints on CR propagation have been inferred assuming a constant (spatially universal and time-independent) and isotropic diffusivity $\kappa_{\rm iso}$, along with an analytic time-independent model of the MW gas distribution that ignores any small-scale phase structure. Most constraints are also based on 'leaky box' or 'flat halo' diffusion models where CRs 'escape' if they go outside a specified volume (historically, a thin disc with height $\sim 200\,$pc). But all these assumptions can be order-of-magnitude incorrect. Small gyro-radii mean diffusion is strongly anisotropic, and MW star formation and ISM structure is strongly time-variable on time-scales well below the CR residence time and spatially variable on scales $\lesssim$kpc. Perhaps most problematic, it is now firmly established that essentially all galaxies are embedded in massive, extended CGM gaseous haloes containing *most* of the baryons, with smooth, shallow density profiles extending to $\gtrsim 200\,$kpc (with scale lengths $\sim$ 20–50 kpc; see e.g. Tumlinson, Peeples & Werk 2017, and references therein). In analytic or idealized numerical 'leaky box' or 'flat halo diffusion' CR transport models when a toy-model 'halo' is added (usually a cylinder of height $H_{\rm halo} \sim$ 1–10 kpc), the inferred $\kappa_{\rm iso}$ increases with $\sim H_{\rm halo}$ (Strong & Moskalenko 2001; Vladimirov et al. 2012; Gaggero et al. 2015; Cummings et al. 2016; Guo, Tian & Jin 2016; Jóhannesson et al. 2016; Korsmeier & Cuoco 2016; Evoli et al. 2017; Amato & Blasi 2018), so this effect alone can increase the 'required' diffusivities by factors of $\sim$100.

Making matters more complicated, recent work has shown the properties of the gaseous halo itself can depend strongly on the $\sim$GeV CR transport (Butsky & Quinn 2018; Ji et al. 2020). Moreover, in physically motivated CR transport models, the local diffusivity is typically a strong function of the local plasma properties (strength of turbulence, magnetic field strength, density, ionization level), which vary by *orders of magnitude* on $\sim$ 0.1–100pc scales within the ISM.

However, several recent breakthroughs have made real progress possible. (1) Recent γ-ray observations (mostly from Fermi) have established strong constraints on $\sim$GeV CRs in a number of nearby galaxies, complementing the classical Solar-neighbourhood constraints on inferred CR grammage, residence times, and energy density. Surprisingly, while the most dense starburst systems observed appear to be proton calorimeters, all 'normal' $\sim L_*$ and dwarf galaxies observed (the MW, Andromeda/M31, SMC, LMC, M33) have robust upper limits or detections indicating that at least $\sim$ 95–99 per cent of the $\sim$GeV CRs must escape *without* hadronic collisions, requiring large diffusivities (Lacki et al. 2011; Tang, Wang & Tam 2014; Griffin, Dai & Thompson 2016; Fu, Xia & Shen 2017; Wojaczyński & Niedźwiecki 2017; Lopez et al. 2018; Wang & Fields 2018). (2) Analytic and numerical work explicitly following transport and scattering of CRs on 'micro-scales' (e.g. Bai et al. 2015, 2019; Lazarian 2016; Holcomb & Spitkovsky 2019; van Marle, Casse & Marcowith 2019), coupled to improved intermediate-scale 'effective fluid' theories (e.g. Zank 2014; Zweibel 2017; Thomas & Pfrommer 2019), has begun to yield more detailed prescriptions for the 'effective' transport coefficients of CRs as a function of local plasma properties (appropriate on scales much larger than the CR gyro-radius, but much smaller than the scales of e.g. ISM phases where these properties change dramatically), for both extrinsic turbulence (ET) and self-confinement (SC) scenarios. (3) Cosmological galaxy simulations can now self-consistently model the time-and-space-dependent phase structure of the ISM together with extended CGM haloes, while explicitly following CR populations (Chan et al. 2019; Butsky & Quinn 2018; Hopkins et al. 2020b; Ji et al. 2020; Su et al. 2020).

In this paper, we synthesize these three advances, to directly constrain proposed micro-physical models of $\sim$GeV CR transport. To properly model observables like grammage, residence time, and γ-ray emission, we need to forward-model CR production and transport self-consistently in cosmological simulations which can actually model the ISM/CGM gaseous haloes and phase structure (since these strongly influence the observables). The Feedback In Realistic Environment (FIRE)[1] simulations we use here have been shown to reproduce MW and dwarf galaxies with CGM phase structure and gas mass profiles (van de Voort et al. 2016; Hafen et al. 2019; Ji et al. 2020; Su et al. 2020), outflow properties (Hopkins, Quataert & Murray 2012b; Hopkins et al. 2013b; Muratov et al. 2015; Hayward & Hopkins 2017), ISM phases and detailed molecular cloud properties (Hopkins, Quataert & Murray 2012a; Guszejnov, Hopkins & Ma 2017; Guszejnov et al. 2020), morphologies (Wheeler et al. 2017; El-Badry et al. 2018a, b; Garrison-Kimmel et al. 2018), star formation histories and masses (Hopkins et al. 2014, 2018b; Garrison-Kimmel et al. 2019), and magnetic field strengths/morphologies (Su et al. 2017, 2018, 2019; Guszejnov et al. 2020), all consistent with state-of-the-art observations. These simulations reach $\sim$pc resolution, which is much larger than the gyro-radii $r_{\rm L}$ of $\sim$GeV CRs, so we cannot a priori predict the CR scattering rates (or diffusivity/streaming speeds). However, this resolution *is* sufficient to begin to resolve two crucial scales: (1) the scales of the dominant ISM/CGM phase structures and driving scales of ISM turbulence and (2) the CR 'mean free path' or deflection length $\lambda_{\rm mfp} \sim c/\nu$ (where ν is the CR scattering rate), for the observationally favoured values of ν. This means that if we have a model for the effective diffusion coefficient or 'streaming speed' of CRs as a function of local plasma properties (or for the more complicated hybrid transport parameters that arise in self-confinement theories), we can self-consistently resolve the full end-to-end CR transport and the observables above on galactic scales. In our previous work (Chan et al. 2019; Hopkins et al. 2020b; Ji et al. 2020; Su et al. 2020), we did this assuming a simplified anisotropic streaming+diffusion model with a constant parallel diffusivity $\kappa_{\parallel}$ and parallel streaming at $v_{\rm st} = v_A$ (the Alfvén speed). These works showed that one can obtain converged solutions that reproduce the observed γ-ray constraints as well as MW grammage/residence-time constraints. We now extend this to a variety of detailed physical models for CR propagation, motivated by both ET and SC models for scattering.

In Section 2, we briefly review the simulation numerical methods, and in Section 3 we review the different micro-physical CR transport models surveyed. Section 4 presents the results and compares to present observational constraints. Section 5 discusses and compares these in more detail, considers which models are ruled out and discusses what missing physics might reconcile these with observational constraints, and compares simple analytic or order-of-magnitude expectations for various quantities. Section 6 briefly compares to historical simulation and analytic models. We summarize in Section 7.

2 METHODS

2.1 Overview and non-CR physics

The simulations here extend those in Chan et al. (2019, Paper I) and Hopkins et al. (2020b, Paper II), where numerical details are described. We only briefly summarize these and the non-CR

[1] http://fire.northwestern.edu

physics here. The simulations are run with GIZMO[2] (Hopkins 2015), in its meshless finite-mass MFM mode (a mesh-free finite-volume Lagrangian Godunov method). All simulations include ideal magnetohydrodynamics (MHD), solved as described in Hopkins (2016) and Hopkins & Raives (2016), and fully anisotropic Spitzer–Braginskii conduction and viscosity (implemented as in Paper II; see also Hopkins 2017; Su et al. 2017). Gravity is solved with adaptive Lagrangian force softening (matching hydrodynamic and force resolution). We treat cooling, star formation, and stellar feedback following the FIRE-2 implementation of the FIRE physics (all details in Hopkins et al. 2018b). We follow 11 abundances (Colbrook et al. 2017; Escala et al. 2018); cooling chemistry from ~ 10–10^{10} K accounting for a range of processes including metal-line, molecular, fine-structure, photoelectric, and photoionization, including local sources and the Faucher-Giguère et al. (2009) meta-galactic background (with self-shielding) and tracking detailed ionization states; and star formation in gas which is dense ($>1000\,{\rm cm^{-3}}$), self-shielding, thermally Jeans-unstable, and locally self-gravitating (Hopkins, Narayanan & Murray 2013a; Grudić et al. 2018). Once formed, stars evolve according to standard stellar evolution models accounting explicitly for the mass, metal, momentum, and energy injection via individual SNe (Ia and II) and O/B or AGB-star mass-loss (for details, see Hopkins et al. 2018a), and radiation (including photoelectric and photoionization heating and radiation pressure with a five-band radiation-hydrodynamic scheme; Hopkins et al. 2020a). Our models are fully cosmological 'zoom-in' simulations, evolving a large box from redshifts $z \gtrsim 100$, with resolution concentrated in a $\sim$ 1–10 Mpc co-moving volume centred on a 'target' halo of interest. While there are many smaller galaxies in that volume, for the sake of clarity we focus just on the properties of the 'primary' (i.e. best-resolved) galaxies in each volume. The galaxies studied are summarized in Table 2.

2.2 CR physics and basic equations

All simulations here also include CRs as described in Papers I and II. We evolve a single-bin ($\sim$GeV) of CRs, or (equivalently) a constant spectral distribution, as a relativistic fluid (energy density $e_{\rm cr}$, pressure $P_{\rm cr} = (\gamma_{\rm cr}-1)\,e_{\rm cr}$ with $\gamma_{\rm cr}=4/3$), with a fixed fraction $\epsilon_{\rm cr} = 0.1$ of the initial SNe ejecta kinetic energy in each explosion injected into CRs. CRs contribute to the total pressure which appears in the gas momentum equation according to the local strong-coupling approximation. Throughout, we denote the CR gyro/Larmor radius $r_{\rm L} \equiv c/\Omega$ with c the speed of light and $\Omega = Z\,e\,c\,|\boldsymbol{B}|/E_{\rm cr}$ the gyro frequency of the CRs (where e is the electron charge and $E_{\rm cr}/Z \equiv \gamma_{\rm L}$ GeV, with $\gamma_{\rm L}\sim 1-10$ for the CR protons of interest here).

Following Papers I and II, CRs then obey a standard energy and flux equation (see e.g. McKenzie & Voelk 1982):

$$\frac{\partial e_{\rm cr}}{\partial t} + \nabla\cdot(\boldsymbol{u}\,h_{\rm cr} + \boldsymbol{F}) = \boldsymbol{u}\cdot\nabla P_{\rm cr} - \Lambda_{\rm st} - \Lambda_{\rm coll} + S_{\rm in} \tag{1}$$

$$-\frac{(\gamma_{\rm cr}-1)}{\kappa_*}\,\boldsymbol{F} = \nabla_\parallel P_{\rm cr} + \frac{\mathbb{D}_t\boldsymbol{F}}{\tilde{c}^2}. \tag{2}$$

In equation (1), $\boldsymbol{u}$ is the gas fluid velocity, $\boldsymbol{F}$ the CR flux in the fluid frame, $h_{\rm cr} \equiv e_{\rm cr} + P_{\rm cr}$ the CR enthalpy, $S_{\rm in}$ the CR source injection, and $\Lambda_{\rm st} = {\rm MIN}(v_A,\,v_{\rm st})\,|\nabla_\parallel P_{\rm cr}|$ represents 'streaming losses', which arise because gyro-resonant Alfvén waves (unresolved wavelengths $\sim r_{\rm L}$) are excited by CR streaming (with speed $v_{\rm st}$, defined below) and rapidly damp (Wentzel 1968; Kulsrud & Pearce 1969). These losses are limited to the Alfvén speed v_A, as we show below (see also Paper I and Ruszkowski et al. 2017). The $\Lambda_{\rm coll}$ term represents collisional (hadronic and Coulomb) losses with $\Lambda_{\rm coll} = 5.8\times10^{-16}\,{\rm s^{-1}\,cm^3}\,(n_{\rm n}+0.28\,n_e)\,e_{\rm cr}$ (with $n_{\rm n}$ and $n_{\rm e}$ the nucleon and free electron number densities), following Guo & Oh (2008). Of $\Lambda_{\rm coll}$, all Coulomb (the $n_{\rm e}$ term) and $\sim$1/6 of the hadronic ($n_{\rm n}$) losses are thermalized; $\Lambda_{\rm st}$ is thermalized as well. In equation (2), $\nabla_\parallel P_{\rm cr} \equiv (\hat{\boldsymbol{b}}\otimes\hat{\boldsymbol{b}})\cdot\nabla P_{\rm cr} = \hat{\boldsymbol{b}}\,(\hat{\boldsymbol{b}}\cdot\nabla P_{\rm cr})$ is the parallel derivative, $\tilde{c}$ is the maximum (physical or numerical) CR free-streaming/signal speed ($\geq 1000\,{\rm km\,s^{-1}}$ here), κ_* is a local effective diffusivity (defined below), and $\mathbb{D}_t\boldsymbol{F} \equiv \hat{\boldsymbol{F}}\,[\partial|\boldsymbol{F}|/\partial t + \nabla\cdot(\boldsymbol{u}\,|\boldsymbol{F}|) + \boldsymbol{F}\cdot\{(\hat{\boldsymbol{F}}\cdot\nabla)\,\boldsymbol{u}\}]$ is the derivative operator derived in Thomas & Pfrommer (2019) from a two-moment expansion of the relativistic Vlasov equation for CRs (assuming a locally gyrotropic CR distribution in the fluid frame and vanishingly small gyro radii, to $\mathcal{O}(v^2/c^2)$).[3] Because the gyro radii of GeV CRs are vastly smaller than resolved scales, they move along the field lines, with $\hat{\boldsymbol{F}} = \hat{\boldsymbol{b}}$ by construction.

As shown in Papers I and II and below, the overwhelmingly dominant uncertainty in CR transport on these scales comes from the form of κ_*, which we will explore extensively. Variations to other choices above, e.g. turning off the sink terms $\Lambda_{\rm st}$ or $\Lambda_{\rm coll}$, otherwise altering the functional form of the flux equation (2) (or simply solving a single energy equation, specifying some equilibrium $\boldsymbol{F}$), varying $\tilde{c}$ widely, or varying $\epsilon_{\rm cr} \sim 0.05$–$0.2$, all have minor or negligible effects on our results. These are reviewed in Appendix D.

2.3 Effective CR 'transport parameters'

We explicitly evolve $\boldsymbol{F}$ according to equation (2). However because the bulk CR flux, by construction, always points along the magnetic field direction ($\hat{\boldsymbol{F}} = \hat{\boldsymbol{v}}_{\rm st} = -\hat{\nabla}_\parallel P_{\rm cr} = \pm\hat{\boldsymbol{b}}$), one can always write the instantaneous flux in terms of an *effective* local scalar diffusion and/or streaming coefficient, i.e.

$$\boldsymbol{F} \equiv -\kappa_{\rm eff}\,\nabla_\parallel e_{\rm cr} \equiv \bar{\boldsymbol{v}}_{\rm st,\,eff}\,h_{\rm cr} \equiv -\bar{\kappa}_\parallel\,\nabla_\parallel e_{\rm cr} + \bar{\boldsymbol{v}}_{\rm st}\,h_{\rm cr}, \tag{3}$$

where $\bar{\boldsymbol{v}}_{\rm st} = -\bar{v}_{\rm st}\,(\nabla_\parallel P_{\rm cr})/|\nabla_\parallel P_{\rm cr}|$ is the streaming velocity, defined to point along the $\boldsymbol{B}$-field down the CR pressure gradient. In other words, we can always simply *define* $\kappa_{\rm eff} \equiv |\boldsymbol{F}|/|\nabla_\parallel e_{\rm cr}|$, or re-write pure-diffusion ($v_{\rm st}=0$) as pure-streaming with $\bar{v}_{\rm st}\rightarrow\bar{\kappa}_\parallel/(\gamma_{\rm cr}\,\ell_{\rm cr})$ (where for convenience we define the parallel CR pressure gradient scale length $\ell_{\rm cr} \equiv P_{\rm cr}/|\nabla_\parallel P_{\rm cr}|$), or vice versa ($\bar{\kappa}_\parallel \rightarrow \gamma_{\rm cr}\,\bar{v}_{\rm st}\,\ell_{\rm cr}$).

In quasi-steady state ($\mathbb{D}_t\boldsymbol{F}\rightarrow 0$), the Newtonian limit ($\tilde{c}$ sufficiently large), on scales large compared to the CR mean free path/time ($\sim\kappa_*/\tilde{c}$), or in the 'pure streaming+diffusion' approximation for the flux ($\mathbb{D}_t\rightarrow 0$), equation (2) gives $\boldsymbol{F}\rightarrow\kappa_*\nabla_\parallel e_{\rm cr}$, so $\kappa_*\rightarrow\kappa_{\rm eff} = \bar{\kappa}_\parallel + \gamma\,v_{\rm st}\,\ell_{\rm cr}$ *exactly*. For this and other physical reasons (see Paper I and Jiang & Oh 2018), we therefore write $\kappa_* = \kappa_\parallel + \gamma_{\rm cr}\,v_{\rm st}\,\ell_{\rm cr}$ in equation (2), where we refer to the coefficients $\kappa_\parallel(\boldsymbol{x},\,t,\,\ldots)$ and $v_{\rm st}(\boldsymbol{x},\,t,\,\ldots)$ as the local 'diffusivity' and 'streaming speed', respectively. But we emphasize that these can be *arbitrary functions* of the local plasma properties and their derivatives, so equation (1) does not necessarily behave like a traditional streaming or diffusion equation.

We will explore variations in the functions $\kappa_\parallel$ and $v_{\rm st}$ below, and we will write and refer to both $\kappa_\parallel$ and $v_{\rm st}$, even though once they are arbitrary functions, their individual values are irrelevant to the CR propagation (only the combined function κ_* is meaningful). Our reason for making this distinction between diffusion

[2] A public version of GIZMO is available at http://www.tapir.caltech.edu/~phopkins/Site/GIZMO.html

[3] As discussed in Appendix D, the operator $\mathbb{D}_t$ in equation (2) is very slightly different from that adopted in Paper II, but the difference enters at $\mathcal{O}(1/\tilde{c}^2)$ and has no effect on our conclusions.

and streaming is largely historical, and we stress that the traditional differences in 'diffusive-like' versus 'streaming-like' behaviour only apply when $\kappa_\parallel$ and/or $v_{\rm st}$ are constants. This is explored further in Appendix B3.

2.4 The Alfvén speed and gyro-resonant wavelengths

Ideal or Braginskii MHD, in which the Alfvén speed is $v_A = v_A^{\rm ideal} \equiv (|\boldsymbol{B}|^2/4\pi\,\rho)^{1/2}$, is an excellent approximation on all *resolved* scales in the simulations here (even when $f_{\rm ion} \ll 1$ in e.g. GMCs),[4] But SC models often refer specifically to the Alfvén speed of gyro-resonant Alfvén waves, which are vastly shorter wavelength (parallel wavenumbers $k_\parallel \sim k_{\rm L} = 2\pi/\lambda_{\rm L} \sim 1/r_{\rm L}$) and therefore can have frequencies much larger than the collision frequency between ions and neutrals in GMCs, and so propagate at the 'ion-Alfvén' speed $v_A^{\rm ion} \equiv (|\boldsymbol{B}|^2/4\pi\,\rho_{\rm i})^{1/2} = f_{\rm ion}^{-1/2}\,v_A^{\rm ideal}$ (Skilling 1975). Such short-branch waves are rapidly damped when $f_{\rm ion} \ll 1$, but the models can account for this. So in general when we refer to v_A, we take $v_A = v_A^{\rm ideal}$, but we explicitly note when we consider $v_A^{\rm ion}$.

Anisotropic viscosity in hot, dilute gas formally modifies the Alfvén speed as well (e.g. Kempski, Quataert & Squire 2019), but the fractional change in Alfvén speed is small for the hot ISM and CGM.

3 DIFFERENT CR TRANSPORT MODELS CONSIDERED

Here, we describe the different CR transport models considered in this paper, summarized in Table 1. For each of these models, we have run a suite of cosmological simulations with at least galaxies **m11i**, **m11f**, **m12i**, chosen because these span a range of masses and, at each mass, show representative effects and scalings of CRs on galaxy dynamics in Papers I and II. An illustration of the galaxies and their properties is shown in Fig. 1.

3.1 Constant-diffusivity models

Lacking a physical model, we can simply *assume* $\kappa_\parallel$ = constant. This is commonly done in empirical models for CR transport, and we explored such models extensively in Papers I and II. For the relatively large diffusion coefficients favoured by observations ($\kappa_\parallel \sim 3\times10^{29-30}\,{\rm cm^2\,s^{-1}}$, see Section 4), we showed in Papers I and II that adding or neglecting an 'additional' CR streaming at trans-Alfvénic or trans-sonic speeds made only a very small difference to our conclusions. This follows from our discussion in Section 2.3: what matters on large scales is not $\kappa_\parallel$ or $v_{\rm st}$ individually but the total transport function $\kappa_* = \kappa_\parallel + \gamma_{\rm cr}\,v_{\rm st}\,\ell_{\rm cr}$, where the second (streaming) term is $\sim 4\times10^{27}\,{\rm cm^2\,s^{-1}}\,(v_{\rm st}/10\,{\rm km\,s^{-1}})\,(\ell_{\rm cr}/{\rm kpc})$. Thus, even factor of $\sim$10 variations in $v_{\rm st}$ around typical trans-Alfvénic values amount to $\sim$ 0.1–10 per cent variations in κ_* (for $\kappa_\parallel \sim 10^{30}\,{\rm cm^2\,s^{-1}}$), compared to the order-of-magnitude variations in $\kappa_* \sim \kappa_\parallel$ which fall within the 'allowed' range.

We stress that these models have no particular physical motivation: they simply provide an empirical reference point for the transport speeds 'needed' (in the ISM and near-field CGM where e.g. γ-ray emission originates) to reproduce observational constraints.

[4]Formally, the ion-neutral 'strong-coupling' approximation (ion-neutral collision times are short compared to resolved time-scales) applies on all simulated scales ($\sim$ pc or larger).

3.1.1 Model variant: 'fast' transport in neutral gas, 'slow' in ionized gas

In SC scenarios, strong ion-neutral damping can produce rapid transport in primarily neutral gas. In Farber et al. (2018), the authors attempt to approximate this effect with a 'two-κ' model, with a constant-but-different diffusivity in neutral and ionized gas.[5] We therefore consider a similar model, parametrized as

$$\kappa_\parallel = 3\times10^{29}\,{\rm cm^2\,s^{-1}}\left(1 - f_{\rm ion} + \frac{f_{\rm ion}}{30}\right) \quad (4)$$

(with $v_{\rm st} = v_A$), so $\kappa_\parallel = 3\times10^{29}$ or $\kappa_\parallel = 10^{28}\,{\rm cm^2\,s^{-1}}$ in neutral or ionized gas, respectively. This is a useful reference model because it allows us to explore whether CR diffusion must be relatively 'fast' in *both* neutral and ionized gas, or just the densest (neutral) gas.

3.1.2 Model variant: pure-advection and Alfvénic/sonic streaming-only

If $\kappa_* \rightarrow 0$ (i.e. $\kappa_\parallel \rightarrow 0$ and $v_{\rm st} \rightarrow 0$), then $\boldsymbol{F} \rightarrow \boldsymbol{0}$ and CRs are purely advected with gas. It is well established that this cannot possibly reproduce observations in the MW and nearby galaxies. If the only CR transport beyond advection were streaming with trans-Alfvénic or trans-sonic speeds, this is identical to our default constant-$\kappa_\parallel$ models with $\kappa_\parallel \rightarrow 0$ (and $v_{\rm st} \sim v_A$). In the MW warm ISM, with $v_A \sim c_s \sim 10\,{\rm km\,s^{-1}}$, this gives effective diffusivities $\kappa_{\rm eff} \sim v_{\rm A}\,\ell_{\rm cr} \sim 10^{27}\,{\rm cm^2\,s^{-1}}$, much lower than our preferred $\kappa_{\rm eff}$. These cases are considered explicitly in Papers I and II, with $v_{\rm st} \sim 0$, v_A, $3\,v_A$, $10\,v_A$, $v_{\rm fast}$, $3\,v_{\rm fast}$ (where $v_{\rm fast}^2 = c_s^2 + v_A^2$ is the fastest ideal-MHD wavespeed), where we showed all produce far too-slow CR transport and overpredict observed γ-ray fluxes from nearby galaxies by $\sim$1–2 dex. So we do not consider these cases further, except as the obvious limit when $\kappa_\parallel \rightarrow 0$.

3.2 Extrinsic turbulence scenarios

The CR diffusivity is $\kappa_{\rm eff} \sim c^2/3\,\nu$, where ν is the scattering rate ($\lambda_{\rm mfp} \sim c/\nu$ is the CR mean free path). In the standard picture, CRs scatter off of magnetic-field fluctuations $\delta\boldsymbol{B}$, with a strong preference for 'resonant' fluctuations $\delta\boldsymbol{B}[r_{\rm L}]$, i.e. fluctuations with parallel wavenumber $k_\parallel \sim k_{\rm L} \sim 1/r_{\rm L}$. Simple quasi-linear theory calculations give the scattering rate $\nu \sim \Omega\,|\delta\boldsymbol{B}[r_{\rm L}]|^2/|\boldsymbol{B}|^2$ (e.g. Jokipii 1966; Wentzel 1968; Skilling 1971).

In the simplest possible 'ET' model (e.g. Jokipii 1966; Voelk 1975), we can estimate $\kappa_{\rm eff}$ by extrapolating $|\delta\boldsymbol{B}[r_{\rm L}]|$ from a turbulent power spectrum with (1D) Alfvén Mach number $\mathcal{M}_A = \mathcal{M}_A[\ell_{\rm turb}] \equiv |\delta\boldsymbol{B}[\ell_{\rm turb}]|/|\boldsymbol{B}| \approx |\delta\boldsymbol{v}[\ell_{\rm turb}]|/v_A^{\rm ideal}$ on some resolved scale $\ell_{\rm turb}$. While very high energy CRs (with large $r_{\rm L}$) may scatter significantly on $\ell_{\rm turb}$ scales directly, we are interested in low-energy CRs with $r_{\rm L} \sim 10^{-6}$ pc. Such scales are smaller than the damping/viscous scale for fast/acoustic modes, while Alfvénic modes, although not strongly damped, are highly anisotropic on these scales, which must be taken into account for estimates of ν (as we do below). None the less, as a reference model, let us assume a Goldreich & Sridhar (1995)-type (GS95) cascade ($E_\parallel \propto k_\parallel^{-2}$), giving

$$\frac{\kappa_\parallel}{c\,r_{\rm L}} \sim \frac{|\boldsymbol{B}|^2}{|\delta\boldsymbol{B}[k_\parallel \sim 1/r_{\rm L}]|^2}\,f_{\rm turb} \sim \mathcal{M}_A^{-2}\,\frac{\ell_{\rm turb}}{r_{\rm L}}\,f_{\rm turb},$$

$$\kappa_\parallel \sim 10^{32}\,{\rm cm^2\,s^{-1}}\,\mathcal{M}_A^{-2}\,\ell_{\rm turb,\,kpc}\,f_{\rm turb}, \quad (5)$$

[5]They adopted $\kappa_\parallel = 10^{29}$ or $3\times10^{27}\,{\rm cm^2\,s^{-1}}$ in gas below/above $T = 10^4$ K, using temperature as a proxy for ionization state.

Table 1. Subset of CR transport models studied. All models include star formation, stellar feedback, MHD, anisotropic conduction, and viscosity.

Name	Description	Ref.	$\langle\kappa_{\rm eff}^{\rm iso}\rangle_{29}^{v}$	L_γ, X_s?	$\langle e_{\rm cr}\rangle$
CD:	**Constant-diffusivity models (Section 3.1, equation 3)**: $\kappa_\parallel = \kappa_{29}\,10^{29}\,{\rm cm^2\,s^{-1}}$, varied $v_{\rm st}\sim v_A$				
$\kappa_{29}=0$	$\kappa_{29}=0$, $v_{\rm st}=(0,\,1,\,3,\,4,\,1+\beta^{1/2},\,3\,[1+\beta^{1/2}])\,v_A$ (Section 3.1.2)	*a*	$\lesssim$0.01	× (high)	40
$\kappa_{29}=0.03$	$\kappa_{29}=0.03$, $v_{\rm st}=(1,\,3)\,v_A$	*a*	0.015	× (high)	50
$\kappa_{29}=0.3$	$\kappa_{29}=0.3$, $v_{\rm st}=(0,\,1,\,3)\,v_A$	*a*	0.1	× (high)	8
$\kappa_{29}=3$	$\kappa_{29}=3$, $v_{\rm st}=(0,\,1,\,3)\,v_A$ (favoured models in Papers I and II)	*a*	1	✓	1
$\kappa_{29}=30$	$\kappa_{29}=30$, $v_{\rm st}=v_A$	*a*	10	✓	0.4
$\kappa_{29}=300$	$\kappa_{29}=300$, $v_{\rm st}=v_A$	*a*	100	○ (low)	0.04
$\kappa_{\rm ion\text{-}neutral}$	$\kappa_{29}=3$ in neutral gas, =0.1 in ionized gas (Section 3.1.1, equation 4)	*b*	0.05	× (high)	20
ET:	**Extrinsic turbulence models (Section 3.2, equation 5)**: $\kappa_\parallel=\mathcal{M}_A^{-2}\,c\,\ell_{\rm turb}\,f_{\rm turb}$, varied $f_{\rm turb}$				
Alfvén-C00	$f_{\rm turb}=0.14\,(c_s/v_A)/\ln(\ell_{\rm turb}/r_{\rm L})$: anisotropic GS95 spectrum of Alfvén modes	*c*	1500	○ (low)	0.2
Alfvén-C00-Vs	As Alfvén-C00, adding additional 'streaming' $v_{\rm st}=v_A$ or $v_A^{\rm ion}$	–	1500	○ (low)	0.2
Alfvén-YL02	$f_{\rm turb}=70\,(c/v_A)^{5/11}\,(\ell_{\rm turb}/r_{\rm L})^{9/11}$: modified non-resonant Alfvén scattering	*d*	$>10^4$	○ (low)	0.001
Alfvén-Hi	$f_{\rm turb}=1000$: arbitrarily changed $f_{\rm turb}$	–	400	○ (low)	0.02
Alfvén-Max	$f_{\rm turb}=1$: GS95 Alfvén scattering ignoring gyro-averaging/anisotropy	–	1	✓	2
Fast-YL04	$f_{\rm turb}=f(\lambda_{\rm damp})$: non-resonant fast modes, damped below $\lambda_{\rm damp}$	*e*	80	○ (low)	0.006
Fast-Max	As YL04, neglect ion-neutral and $\beta>1$ viscous damping	*e*	6	✓	1
Fast-Mod	$f_{\rm turb}\sim 1000\times$ the 'Fast-Max' value (different spectrum, broadening)	–	700	○ (low)	0.04
Fast-NoDamp	$f_{\rm turb}=(r_{\rm L}/\ell_{\rm turb})^{1/2}$: Fast-YL04, ignoring any fast-mode damping	–	0.003	× (high)	3
Fast-NoCDamp	$f_{\rm turb}$ given by Fast-Max with viscous damping only	–	0.03	× (high)	5
Iso-K41	$f_{\rm turb}=(r_{\rm L}/\ell_{\rm turb})^{1/3}$: isotropic, undamped K41 cascade down to $<r_{\rm L}$	*f*	0.004	× (high)	0.4
Fast-Max+Vs	As Fast-YL04, adding additional 'streaming' $v_{\rm st}=v_A$ or $v_A^{\rm ion}$	–	7	✓	1
SC:	**Self-confinement models (Section 3.3, equation 7)**: $\kappa_\parallel\propto\Gamma$ (damping), $v_{\rm st}=v_A^{\rm ion}$, varied Γ				
Default	Default scalings for $\Gamma=\Gamma_{\rm in}+\Gamma_{\rm turb}+\Gamma_{\rm LL}+\Gamma_{\rm NLL}$, Appendix A	–	0.02	× (high)	10
Non-Eqm	Replace $\kappa_\parallel$, $v_{\rm st}$ with evolved gyro-resonant $\delta\boldsymbol{B}[r_{\rm L}]$ (Section 3.3.2)	–	0.03	× (high)	4
10 GeV	Adopt $\gamma_{\rm L}=10$ instead of =1 (typical $E_{\rm cr}/Z\sim 10$ GeV; Section 3.3.3)	–	0.03	× (high)	15
$v_A^{\rm ideal}$	Adopt $v_A=v_A^{\rm ideal}$ instead of $v_A^{\rm ion}$ in equation (7; Section 3.3.1)	–	0.007	× (high)	15
$f_{\rm QLT}-6$	Multiply $\kappa_\parallel$ in equation (7) by $f_{\rm QLT}$ (weaker growth or stronger damping; Section 3.3.4)	–	0.05	× (high)	10
$f_{\rm QLT}-6$, 10 GeV	Combines '$f_{\rm QLT}-6$' and '10 GeV' models	–	0.1	× (high)	8
$f_{\rm QLT}-6$, $v_A^{\rm ideal}$	Combines '$f_{\rm QLT}-6$' and '$v_A^{\rm ideal}$' models	–	0.04	× (high)	10
$f_{\rm QLT}-100$	Multiply $\kappa_\parallel$ in equation (7) by $f_{\rm QLT}=100$	–	5	✓	0.3
$f_{\rm cas}-5$	$f_{\rm cas}=5$ in $\Gamma_{\rm turb}$ and $\Gamma_{\rm LL}$	–	0.06	× (high)	8
$f_{\rm cas}-50$	$f_{\rm cas}=50$ in $\Gamma_{\rm turb}$ and $\Gamma_{\rm LL}$	–	2	✓	0.3
$f_{\rm cas}-500$	$f_{\rm cas}=500$	–	10	✓	0.4
$f_{\rm cas}-$DA	$f_{\rm cas}=(\ell_{\rm turb}/r_{\rm L})^{1/10}$, for a 'dynamically aligned' perpendicular spectrum ($\sim k_\perp^{-3/2}$)	–	0.02	× (high)	10
$f_{\rm cas}-$B73	$f_{\rm cas}={\rm MIN}(1,\,\mathcal{M}_A^{-1/2})$, for a B73 spectrum above ℓ_A	–	0.005	× (high)	20
$f_{\rm cas}-$L16	$f_{\rm cas}$ follows a multicomponent cascade model from L16	*g*	0.004	× (high)	15
$f_{\rm cas}-$K41	$f_{\rm cas}=\mathcal{M}_A^{-1/2}\,(\ell_{\rm turb}/r_{\rm L})^{1/6}$ for an isotropic, undamped K41 cascade	–	15	✓	0.3
NE,$f_{\rm cas}-$L16	As 'Non-Eqm' but with $f_{\rm cas}$ following $f_{\rm cas}-$L16 model	–	0.01	× (high)	4
NE,$f_{\rm QLT}-100$	As 'Non-Eqm' but with $f_{\rm QLT}=100$	–	7	✓	0.3
ET+SC:	**Combined extrinsic turbulence and self-confinement (Section 3.4)**: $\nu_{\rm total}=\sum\nu_i$ (sum ET+SC terms), $v_{\rm st}=v_A^{\rm ion}$				
A+F+SC100	ET:Alfvén-C00 + ET:Fast-Max + SC:$f_{\rm turb}=100$	–	2	✓	1
A+SC100	ET:Alfvén-C00 + SC:$f_{\rm turb}=100$	–	5	✓	0.3

Note. Summary of the different CR transport models (models for the effective transport coefficients $\kappa_\parallel$ and $v_{\rm st}$ in equation 2). Column include: (1) Name. (2) Description. (3) References where previously studied. (4) $\langle\kappa_{\rm eff}^{\rm iso}\rangle_{29}^{v}$: time (redshifts $z<0.1$, sampled each ~ 10 Myr) and space (galactocentric radii <10 kpc) and angle (isotropic-equivalent) averaged, *scattering-rate-weighted* effective diffusivity $\kappa_{\rm eff}^{\rm iso}\equiv|\boldsymbol{F}_{\rm cr}|/|\nabla e_{\rm cr}|$ (in units of $10^{29}\,{\rm cm^2\,s^{-1}}$) in our MW-like (**m12i**) simulations. (5) L_γ, X_s: qualitative comparison of the predicted γ-ray luminosity and MW grammage to observational constraints, for dwarf (**m11i**), intermediate (**m11f**), and MW-mass (**m12i**) galaxies. A ✓ indicates consistency with observations, 'high' or 'low' indicates the prediction is too high or low. (6) $\langle e_{\rm cr}\rangle$, the time-and-space averaged, *volume-weighted* mean CR energy density (in ${\rm eV\,cm^{-3}}$) in our MW-like (**m12i**) simulations at $z<0.1$ at approximately the solar position (averaged in the thin disc in a galactocentric radial annulus from 7–9 kpc with height ±250 pc). Models are grouped by categories (labelled). Models in red produce excessive confinement and are ruled out by γ-ray observations and MW constraints. Models in cyan produce less confinement than observed: these are allowed, but cannot dominate scattering. Models in black produce reasonable agreement with the observations. References: (*a*) Paper I, (*b*) Farber et al. (2018), (*c*) Chandran (2000), (*d*) Yan & Lazarian (2002), (*e*) Yan & Lazarian (2004, 2008), (*f*) Jokipii (1966), (*g*) Lazarian (2016). Different turbulent power spectra include GS95 (Goldreich & Sridhar 1995), K41 (Kolmogorov 1941), 'dynamically aligned' (Boldyrev 2006), B73 (Burgers 1973).

where $\ell_{\rm turb,\,kpc}\equiv\ell_{\rm turb}/{\rm kpc}$, and we absorb all the microphysics of turbulence and scattering into $f_{\rm turb}$.

3.2.1 *Model variant: turbulent structure assumptions*

There is an extensive literature regarding the 'correct' form of equation (5) (or, equivalently, $f_{\rm turb}$) for ET (see e.g. Zweibel 2013, and references therein). We cannot possibly be comprehensive here, so we focus on a few models chosen to bracket a range of possibilities. Note that the expressions proposed for $f_{\rm turb}$ or κ are often very complicated: we simplify these to order-of-magnitude scalings for the parameter space of interest ($\sim$ GeV CRs, etc.).

(i) **Alfvén-C00**: Chandran (2000) attempt to self-consistently derive κ_* in a Goldreich & Sridhar (1995) cascade, accounting for anisotropy. For all limits relevant here, their result (equation 16

Table 2. Zoom-in simulation volumes (details in Paper II). All units are physical.

Simulation name	$M_{\rm halo}^{\rm vir}$ ($M_\odot$)	$M_*^{\rm (NoCR)}$ ($M_\odot$)	$M_*^{\rm (CR)}$ ($M_\odot$)	$m_{i,\,1000}$ ($1000\,M_\odot$)	$\langle\epsilon_{\rm gas}\rangle^{\rm sf}$ (pc)	Notes
m11i	6.8e10	6e8	(2–7)e8	7.0	1.3	Dwarf galaxy ($\sim$ SMC-mass), with episodic 'bursty' star formation
m11f	5.2e11	4.0e10	(1.5–4)e10	12	1.8	Late-type galaxy, with intermediate surface densities
m12i	1.2e12	7.0e10	(2.5–8)e10	7.0	1.4	$\sim L_*$ galaxy in a 'massive' halo, dense CGM and higher surface density

Note. Properties of the 'primary' galaxy in each zoom-in volume at $z = 0$, including: virial mass ($M_{\rm halo}^{\rm vir}$), stellar mass M_* in the our reference 'no CRs' run ($M_*^{\rm (NoCR)}$) from Paper II, and full range of stellar masses in our runs here with CRs but different transport physics ($M_*^{\rm (CR)}$), mass resolution ($m_{i,\,1000}$), Plummer-equivalent force softening at the mean density of star formation ($\langle\epsilon_{\rm gas}\rangle^{\rm sf}$; note the actual softening is adaptive and varies accordingly).

therein) gives $f_{\rm turb} \approx (0.14/\ln(\ell_{\rm turb}/r_{\rm L}))\,(c/v_A) \sim 1000\,n_1^{1/2}\,B_{\mu\rm G}^{-1}$ (where $n_1 = \rho/(m_p\,{\rm cm}^{-3})$). Here, $f_{\rm turb} \gg 1$ arises because the GS95 cascade has power at $k_{\rm L}$ only for $k_\perp \gg k_\parallel$, which leads to an effective 'reduction factor' in scattering from gyro-averaging.

(ii) **Alfvén-YL02**: Yan & Lazarian (2002) dismiss the dominant non-resonant pitch-angle scattering term from Chandran (2000) as spurious, and argue that one should include only the much weaker resonant scattering term (equation 17 in Chandran 2000), modified slightly by the factor $\sim \Gamma[13/2,\,(\ell_{\rm turb}/r_{\rm L})^{1/3}\,(c_s/c)^{2/3}]$ owing to their different assumed form of the cross-correlation tensor (equation 8 in Yan & Lazarian 2002). This gives $f_{\rm turb} \sim 7\times10^{-4}\,(c/v_A)^{5/11}\,(\ell_{\rm turb}/r_{\rm L})^{9/11} \sim 3\times10^6\,n_1^{0.2}\,B_{\mu\rm G}^{0.4}\,\ell_{\rm turb}^{0.8}\,\gamma_{\rm L}^{-0.8}\,\mathcal{M}_A^{-2.5}$. This is so large that it produces totally negligible confinement/scattering.

(iii) **Fast-YL04**: Yan & Lazarian (2004, 2008) argue that fast magnetosonic modes could dominate CR scattering despite most mode angles $\hat{\mathbf{k}}$ being strongly damped below wavelengths $\lambda_{\rm damp} \gg r_{\rm L}$, if (1) they are isotropic with a shallow power spectrum, (2) non-resonance broadening enhances transit-time damping (TTD), and (3) gyro-resonant ($k \approx k_\parallel \approx 1/r_{\rm L}$) parallel fast modes with $\hat{\boldsymbol{k}} \approx \hat{\boldsymbol{b}}$ are undamped. Using their assumptions (see Appendix C), $\lambda_{\rm damp}$ is then set by the maximum of either collisionless (Landau) or viscous damping: when collisionless dominates we can approximate $f_{\rm turb} \sim 2\,(\pi\,m_e\,\beta/4\,m_p)^{1/2} \sim 0.04\,\beta^{1/2}$, and when viscous dominates we have $f_{\rm turb} \sim \mathcal{M}_A^{5/3}\,{\rm Re}^{-1/3}\,(\ell_{\rm turb}/r_{\rm L})^{1/6}$, where ${\rm Re} \equiv (\mathcal{M}_A\,v_A\,\ell_{\rm turb})/\nu_{\rm v}$ is the Reynolds number with $\nu_{\rm v}$ the kinematic viscosity.[6] However, even given these assumptions, efficient confinement by fast modes requires near fully ionized gas ($f_{\rm neutral} \ll f_{\rm n},\ 0 \approx 0.001\,(n_1\,\beta)^{-3/4}\,T_4^{1/4}\,(\ell_{\rm turb,\,kpc}\,\gamma_{\rm L})^{-1/2}$) *and* low $\beta < 1$, otherwise damping of the gyro-resonant fast modes gives extremely large κ.[7] We approximate these 'cut-offs' by multiplying $f_{\rm turb}$ by a factor $f_{\rm cut} = \exp\{(f_{\rm neutral}/f_{\rm n},\ 0)^4 + (\beta/0.1)^{1.5}\}$ (see Appendix C).

(iv) **Fast-Max**: If we make the ad hoc assumption that some other physics contributes large scattering rates at small pitch angles, or simply neglect any damping of gyro-resonant parallel fast modes, then we approximately obtain the 'Fast-YL04' model but without the 'cut-off' terms suppressing scattering where $f_{\rm neutral} \gtrsim 10^{-3}$ or $\beta \gtrsim 1$. We consider this model ($f_{\rm cut} = 1$) for the sake of reference, if the fast-mode scattering rates for well-ionized, low-β gas were simply applied everywhere in the ISM.

[6]We take $\nu_{\rm v} \sim 10^{18}\,{\rm cm^2\,s^{-1}}\,T_4^{1/2}\,\rho_{-24}^{-1}\,(0.6\,f_{\rm ion}\,T_4^2 + 300\,f_{\rm neutral})$ to be the sum of Braginskii (dominant in ionized gas) and atomic collisional (dominant in neutral gas) viscosities (Spitzer & Härm 1953). To interpolate between collisionless/viscous regimes we simply take the maximum $f_{\rm turb}$ defined by either.

[7]See e.g. Yan & Lazarian (2004), who show that any models with $\beta \geq 1$, such as their 'hot ionized medium' (HIM) model, or with non-negligible neutrals, such as their warm neutral (WNM) or cold cloud (CNM or DC) models, give $\kappa_\parallel \gg 10^{33}\,{\rm cm^2\,s^{-1}}$.

(v) **Fast-Mod**: Yan & Lazarian (2004, 2008) make a number of uncertain assumptions in deriving the effect of fast modes. For example, they assume a fast-mode spectrum $\propto k^{-3/2}$, but the simulations in Cho & Lazarian (2003) used to justify this choice are in several cases more consistent with Kolmogorov (1941) (K41; $k^{-5/3}$) or even Burgers (1973) (B73; k^{-2}) spectra (as others have argued for fast modes in the ISM, e.g. Boldyrev, Nordlund & Padoan 2002; Schmidt, Federrath & Klessen 2008; Kritsuk et al. 2007; Burkhart et al. 2009; Pan, Padoan & Kritsuk 2009; Hopkins 2013), the latter of which would give $f_{\rm turb} \sim 1$. They also assume the non-linear TTD terms are 'broadened' with the maximum possible broadening (given by the driving-scale $\delta\boldsymbol{B}/|\boldsymbol{B}|$, despite $r_{\rm L} \ll \lambda_{\rm damp} \ll \ell_{\rm turb}$); modifying this would increase $f_{\rm turb}$ by a large (exponential) factor (Voelk 1975). Lacking a more detailed model, we consider a case with $f_{\rm turb}$ equal to the 'Fast-Max' model times 1000.

(vi) **Iso-K41**: If we *entirely* ignore anisotropy and damping, and extrapolate an isotropic Kolmogorov (1941) spectrum from $\ell_{\rm turb}$ to $r_{\rm L}$, we obtain $f_{\rm turb} \sim (r_{\rm L}/\ell_{\rm turb})^{1/3} \sim 0.001\,(\gamma_{\rm L}/B_{\mu\rm G}\,\ell_{\rm turb,\,kpc})^{1/3}$. This model is not physically motivated, since the anisotropy of magnetized turbulence is well understood and observed in the solar wind (Chen 2016), but it provides a useful reference.

We have also run a number of additional variations to gain further insight: (vii) assuming fixed $f_{\rm turb} = 1$ (i.e. assume a GS95 cascade, but *ignore* the effect of anisotropy on scattering calculated by Chandran (2000) and Yan & Lazarian (2002)); (viii) fixed $f_{\rm turb} = 1000$ (not motivated by a specific model, but for reference); (ix) variations of model 'Fast-YL04' neglecting all damping (even more extreme than 'Iso-K41'), so $f_{\rm turb} \sim (r_{\rm L}/\ell_{\rm turb})^{1/2} \sim 10^{-5.5}\,(\gamma_{\rm L}/B_{\mu\rm G}\,\ell_{\rm turb,\,kpc})^{1/2}$; (x) variation of 'Fast-YL04'/'Fast-Max' neglecting all but collisionless damping (similar to 'Iso-K41'); (xi) several variants of 'Iso-K41' as proposed in the literature, e.g. that in Snodin et al. (2016) which gives $f_{\rm turb} \sim 0.003 + 0.3\,(r_{\rm L}/\ell_{\rm turb})^{1/3}$; (xii) versions of models (i)-(v) with an additional streaming with both $v_{\rm st} = v_A^{\rm ideal}$ and $v_A^{\rm ion}$; (xiii) versions of (i)–(v) where we assume a Kolmogorov (1941) or Burgers (1973) spectrum on large (simulation-resolved) scales of $\mathcal{M}_A > 1$, down to the scale ℓ_A where $\mathcal{M}_A[\ell_A] = 1$, then the specified spectrum below this scale (as opposed to a single spectrum on all scales), which modifies $f_{\rm turb}$ by, at most, one power of $\mathcal{M}_A[\ell_{\rm turb}] \sim 1$.

Note that in all of the models in this section except 'Fast-YL04', we neglect ion-neutral damping/ambipolar diffusion in gas with $f_{\rm ion} \ll 1$, which will suppress scattering (increasing $f_{\rm turb}$) substantially in molecular clouds. However, we do consider 'fast transport in neutral gas' elsewhere, and in some of the variants here.

3.3 Self-confinement scenarios

In the SC picture, $|\delta\boldsymbol{B}[r_{\rm L}]|$ is dominated by fluctuations from plasma instabilities self-excited by the CR flux. CRs stream down their number density/pressure gradient with speed $\bar{v}_{\rm st}$, but this excites

gyro-resonant Alfvén waves ($k_\parallel \sim k_L$) with growth rate $\Gamma_{\rm grow} \sim \Omega\,(\gamma_L\,n_{\rm cr}/n_i)\,(\bar{v}_{\rm st}/v_A - 1) \sim v_A\,[|\boldsymbol{F}| - v_A\,h_{\rm cr}]/(e_B\,c\,r_L)$,[8] which in turn scatter the CRs (suppressing $\boldsymbol{F}$). A local quasi-steady state arises in which this growth is balanced by damping of these gyro-resonant waves with rate $\Gamma_{\rm damp}$, giving $\Gamma_{\rm grow} \approx \Gamma_{\rm damp}$ or $|\boldsymbol{F}| - v_A\,h_{\rm cr} = \kappa_\parallel\,|\nabla_\parallel e_{\rm cr}| \sim \Gamma_{\rm damp}\,(e_B\,c\,r_L/v_A)$, i.e. CR transport with

$$\frac{\kappa_\parallel}{c\,r_L} \approx \frac{16}{3\pi}\left(\frac{\ell_{\rm cr}\,\Gamma_{\rm eff}}{v_A}\right)\left(\frac{e_B}{e_{\rm cr}}\right) f_{\rm QLT}, \qquad v_{\rm st} \approx v_A$$

$$\kappa_\parallel \sim 6\times10^{26}\,{\rm cm^2\,s^{-1}}\,\frac{\gamma_L\,\Gamma_{-11}\,\ell_{\rm cr,\,kpc}\,f_{\rm ion}^{1/2}\,n_1^{1/2}\,f_{\rm QLT}}{e_{\rm cr,\,eV}}, \tag{7}$$

where $e_B \equiv |\boldsymbol{B}|^2/8\pi$ is the magnetic energy density, $f_{\rm QLT}$ is a factor we insert to parametrize any deviations from the quasi-linear derivation above, and $\Gamma_{\rm eff} \approx \Gamma_{\rm in} + \Gamma_{\rm turb} + \Gamma_{\rm LL} + \langle\Gamma_{\rm NLL}\rangle + \Gamma_{\rm other}$ represents the damping rate of gyro-resonant Alfvén waves (i.e. $\partial|\delta\boldsymbol{B}|^2/\partial t \sim -\Gamma_{\rm eff}\,|\delta\boldsymbol{B}|^2$), here de-composed into ion-neutral ($\Gamma_{\rm in}$), turbulent ($\Gamma_{\rm turb}$), linear Landau ($\Gamma_{\rm LL}$), non-linear Landau ($\Gamma_{\rm NLL}$), and 'other' ($\Gamma_{\rm other}$) terms (see e.g. Skilling 1971; Holman, Ionson & Scott 1979; Kulsrud 2005; Yan & Lazarian 2008; Enßlin et al. 2011; Wiener, Oh & Guo 2013a; Wiener, Pfrommer & Oh 2017). A derivation of equation (7) is given in Appendix B, and expressions for each of the Γ are given in Appendix A. In the latter equality, $\ell_{\rm cr,\,kpc} \equiv \ell_{\rm cr}/{\rm kpc}$, $e_{\rm cr,\,eV} \equiv e_{\rm cr}/{\rm eV\,cm^{-3}}$, $\Gamma_{-11} \equiv \Gamma_{\rm eff}/10^{-11}\,s^{-1}$. Per Section 2.3, we can combine the streaming+diffusion terms into a 'pure streaming' expression[9] with $v_{\rm st} \rightarrow \bar{v}_{\rm st} = v_A + \kappa_\parallel/(\gamma_{\rm cr}\,\ell_{\rm cr})$:

$$\bar{v}_{\rm st} \rightarrow v_A\left[1 + \frac{4\,c\,r_L\,\Gamma_{\rm eff}\,e_B\,f_{\rm QLT}}{\pi\,v_A^2\,e_{\rm cr}}\right]$$

$$\sim v_A\left[1 + \frac{0.4\,\gamma_L\,\Gamma_{-11}\,f_{\rm ion}\,n_1\,f_{\rm QLT}}{B_{\mu\rm G}\,e_{\rm cr,\,eV}}\right]. \tag{9}$$

Now our uncertainty in κ_* is encapsulated in the damping rates Γ.

We stress that although we can (per Section 2.3) write the CR transport equations in terms of 'diffusion+streaming' coefficients (equation 7) or 'pure (super-Alfvénic) streaming' (equation 9), the behaviour of equations (7)–(9) is distinct from either a traditional 'pure diffusion' (constant-κ) or 'pure-streaming' (constant-$v_{\rm st}$) equation, because the coefficients themselves depend on $e_{\rm cr}$ and its gradient (see Appendix B3).

3.3.1 Model variant: choice of Alfvén speed

The Alfvén speed of interest in equations (7)–(9) is that of the gyro-resonant modes, which as noted in Section 2.3 should naively follow the *ion* Alfvén speed $v_A^{\rm ion} = f_{\rm ion}^{-1/2}\,v_A^{\rm ideal}$ in partially neutral gas. In our 'default' SC model, we therefore adopt $v_A = v_A^{\rm ion}$ in equation (7) (consistency requires the same v_A appear in the 'streaming loss' term $\Lambda_{\rm st} = v_A\,|\nabla_\parallel P_{\rm cr}|$). But while the gyro-resonant wave frequencies are un-ambiguously larger than ion-neutral collision frequencies in GMCs, other aspects of the assumptions used to derive equations (7)–(9) (e.g. how to treat gas advection terms and boosts to/from the frame of the fluid, and how CRs enter the gas momentum equation) implicitly assume the 'gas frame' and 'magnetic-field frame' are the same (which is true on large scales even in GMCs, but breaks down at the gyro-resonant scales if $v_A^{\rm ion} \gg v_A^{\rm ideal}$). Also other time-scales (like the CR travel and scattering times) are much longer than ion-neutral collision times. At a fundamental level, knowing how different terms are modified in this limit requires re-deriving CR fluid models such as Thomas & Pfrommer (2019) for a three-fluid (CR, ion, neutral) system. Lacking this, we simply compare model variants where we assume ideal MHD scalings, so $v_A = v_A^{\rm ideal}$ in equation (7) and $\Lambda_{\rm st}$.

3.3.2 Model variant: non-equilibrium description

Recently, Zweibel (2017) and Thomas & Pfrommer (2019) attempted to derive non-equilibrium 'macroscopic' dynamical equations for $|\delta\boldsymbol{B}[r_L]|$, κ, and $v_{\rm st}$, accounting for un-resolved gyro-resonant waves by explicitly evolving a sub-grid energy density ($e_{A\pm} \sim |\delta\boldsymbol{B}[r_L]|^2/4\pi$) or wave spectrum propagating in the $\pm\hat{\boldsymbol{b}}$ directions. We have implemented the full set of equations from Thomas & Pfrommer (2019) and compare it to our default 'local equilibrium' assumption here. Appendix B details the complete set of modifications to our default equations, but the important difference is that κ_* is replaced with the *explicitly evolved* diffusivities $\kappa_\pm/(c\,r_L) \approx (16/9\pi)\,(e_B/e_{A\pm}) \sim (1/3)\,|\boldsymbol{B}|^2/|\delta\boldsymbol{B}[r_L]|^2$, and the scattering term $\mathbf{F}/3\kappa_*$ becomes $\boldsymbol{g}_+ + \boldsymbol{g}_-$ in the CR flux equation (equation 2). The Alfvén-wave energy densities evolve as $\partial e_{A\pm}/\partial t = \pm\boldsymbol{v}_A\cdot\boldsymbol{g}_\pm - \Gamma_{\rm eff}\,e_{A\pm}$, where $\boldsymbol{g}_\pm \equiv (\boldsymbol{F} \mp \boldsymbol{v}_A\,h_{\rm cr})/3\,\kappa_\pm$ and $\mathbf{v}_A\cdot\boldsymbol{g}_\pm$ represents growth from the gyro-resonant instability. In Appendix B, we show that when the Alfvén energy subsystem reaches local steady-state ($\partial e_{A\pm}/\partial t \rightarrow 0$), which occurs on short time-scales $\sim\Gamma^{-1}$, the non-equilibrium system reduces to our default CR evolution equations, with $\kappa_\parallel$ and $v_{\rm st}$ following equation (7).

3.3.3 Model variant: CR energy

We can also vary the effective CR energy γ_L (=1 GeV in our default) assumed in our single-bin approximation. This should represent an effective energy containing most of the CR pressure, but that could vary between ∼0.5 and 10 GeV, in principle, given present observational and theoretical constraints. We have run several variants assuming $\gamma_L = 0.1$ or 10. However, note that given the damping rates in Appendix A, κ and $v_{\rm st}$ are either independent of γ_L (depending only on $e_{\rm cr}$), or scale as $\gamma_L^{1/2}$ at most. Thus, even order-of-magnitude variation in γ_L produces only factor ∼2–3 differences in $\kappa_{\rm eff}$.

3.3.4 Model variant: different growth or scattering rates

In deriving equation (7) (see also Appendix B), if we either (a) multiply the gyro-resonant Alfvén-wave damping rates $\Gamma_{\rm eff}$ by a factor f; (b) divide the effective scattering rate ν for a given $|\delta\boldsymbol{B}[r_L]|$ by f (or equivalently multiply the time-scale for those waves to isotropize the CR distribution function by f); or (c) divide the

[8] Crudely, the Kulsrud & Pearce (1969) gyro-resonant streaming instability has linear-theory growth rate:

$$\Gamma_{\rm grow} \sim \Omega\left(\frac{\gamma_L\,n_{\rm cr}}{n_i}\right)\left(\frac{\bar{v}_{\rm st}}{v_A} - 1\right) \sim \Omega\left(\frac{e_{\rm cr}}{m_p\,c^2}\right)\left(\frac{m_p}{\rho}\right)\left(\frac{|\boldsymbol{F}| - v_A\,h_{\rm cr}}{v_A\,e_{\rm cr}}\right)$$

$$\sim \Omega\left(\frac{v_A}{c}\right)\left(\frac{|\boldsymbol{F}| - v_A\,h_{\rm cr}}{e_B\,c}\right) \sim v_A\left(\frac{|\boldsymbol{F}| - v_A\,h_{\rm cr}}{e_B\,c\,r_L}\right), \tag{6}$$

using $e_{\rm cr} \sim n_{\rm cr}\,\gamma_L\,m_p\,c^2$ with $\rho \sim n_i\,m_p$, $e_B \sim \rho\,v_A^2$, and $\bar{v}_{\rm st} \sim |\boldsymbol{F}|/h_{\rm cr}$.

[9] It is also common to see equation (9) written in the form

$$\bar{v}_{\rm st} \rightarrow v_A\left[1 + \frac{4\,c\,r_L\,\Gamma_{\rm eff}\,e_B}{\pi\,v_A^2\,e_{\rm cr}}\right] = v_A\left[1 + \frac{2}{\gamma_L\,\pi}\,\frac{\Gamma_{\rm eff}}{\Omega}\,\frac{n_{\rm ion}}{n_{\rm cr}}\right], \tag{8}$$

where $e_{\rm cr} \equiv \gamma_L\,\mu\,n_{\rm cr}\,c^2$, $\rho_{\rm ion} = \mu\,n_{\rm ion}$, $n_{\rm ion}$ and $n_{\rm cr}$ are the ion and CR number densities. This form is less useful for our purposes, however.

growth rate of the gyro-resonant modes $\Gamma_{\rm grow}$ by f, then $\kappa_\parallel$ in equation (7) is multiplied by f. We call this 'fudge factor' $f_{\rm QLT}$, which could have its physical origins in any (or a combination) of the aforementioned effects. Lacking any particular model for $f_{\rm QLT}$, we have simply run simulations with $f_{\rm QLT} = 1,\ 6,\ 100,\ 1000$ (=1 is our default).

3.3.5 *Model variant: turbulent cascade assumptions*

While there is relatively little ambiguity in the ion-neutral damping rate $\Gamma_{\rm in}$, and we will show the non-linear Landau damping $\Gamma_{\rm NLL}$ only dominates in the ISM in models which are excluded by observations, both the 'turbulent' ($\Gamma_{\rm turb}$) and 'linear Landau' ($\Gamma_{\rm LL}$) damping rates scale with the turbulent dissipation/cascade time-scale $t_{\rm cas}$ at wavelengths $\sim r_{\rm L}$, which is not well constrained. In Appendix A, we detail the default model, which, following Farmer & Goldreich (2004), assumes a K41 cascade on super-Alfvénic scales and a GS95 cascade on scales $<\ell_A$ (ℓ_A is the Alfvén scale where $\delta v_{\rm turb}(\ell_A) \sim v_A$). This gives $\Gamma_{\rm turb} = v_A^{\rm ideal}/(r_{\rm L}\,\ell_A)^{1/2}\,f_{\rm cas}$ (with $\Gamma_{\rm LL} \approx 0.4\,\beta\,\Gamma_{\rm turb}$ scaling proportionally), where $f_{\rm cas} = 1$ for these default assumptions. However, if we consider different cascade models, we obtain correspondingly different $f_{\rm cas}$; moreover, the exact damping rates will depend on the specific temporal and spatial structure of the turbulent field on these micro-scales, so any analytic model for $\Gamma_{\rm turb}$ is an order-of-magnitude average estimate (where $f_{\rm cas}$ parametrizes our ignorance).

Our default model assumes $f_{\rm cas} = 1$. We consider several variant assumptions, including (1–3) arbitrarily increasing $f_{\rm cas} = 5,\ 50,\ 500$; (4) assuming a supersonic Burgers (1973) spectrum at scales $>\ell_A$ instead of K41, giving $f_{\rm cas} = {\rm MIN}(1,\ \mathcal{M}_A^{-1/2})$; (5) assuming a 'dynamically aligned' $\sim k^{-3/2}$ spectrum (Boldyrev 2006; see also Iroshnikov 1963; Kraichnan 1965) instead of GS95 below ℓ_A, giving $f_{\rm cas} = (\ell_{\rm turb}/r_{\rm L})^{1/10}$; (6) assuming a pure (isotropic) K41 cascade from the driving scale to $r_{\rm L}$, giving $f_{\rm cas} \approx \mathcal{M}_A^{-1/2}\,(\ell_{\rm turb}/r_{\rm L})^{1/6}$ (this is not well motivated but provides a useful 'upper limit'); and (7) assuming the multicomponent cascade model from Lazarian (2016) which adopts isotropic K41 for $\ell > \ell_A$ with a transition between a 'weak' cascade with form following Montgomery & Turner (1981), Sridhar & Goldreich (1994) on large scales to a GS95 cascade on smaller scales, giving $f_{\rm cas} = {\rm MIN}[\mathcal{M}_A^{1/2},\ \mathcal{M}_A^{7/6}\,(\ell_{\rm turb}/r_{\rm L})^{1/6}]$ when $\mathcal{M}_A < 1$ and $f_{\rm cas} = {\rm MIN}[1,\ \mathcal{M}_A^{-1/2}\,(\ell_{\rm turb}/r_{\rm L})^{1/6}]$ when $\mathcal{M}_A \geq 1$.

3.4 Combined extrinsic turbulence and self-confinement models

Scattering by self-excited and extrinsic fluctuations are not mutually exclusive. Their non-linear interplay is poorly understood, but in quasi-linear theory the scattering rates should add linearly (see Zweibel 2017), giving $\kappa_\parallel^{-1} \sim \kappa_{\rm self}^{-1} + \kappa_{\rm extrinsic}^{-1}$. We have therefore also run simulations adopting $v_{\rm st} = v_A$, $\kappa_\parallel^{-1} = \kappa_{\parallel,\,\rm self}^{-1} + \kappa_{\parallel,\,\rm turb}^{-1}$ where $\kappa_{\parallel,\,\rm self}$ follows equation (7) and $\kappa_{\parallel,\,\rm turb}$ follows equation (5), with several combinations of the 'variant' model assumptions. Usually, one model (typically the ET model) has much-larger κ (much lower scattering rate), so the prediction simply becomes identical to that of the model with the lower κ (higher ν). Even in the rare cases where the two contribute comparably (e.g. using 'Fast-Max' for $f_{\rm turb}$ and $f_{\rm cas} = 500$), this simply gives similar behaviour to both 'individual' models and so does not change any of our conclusions regarding which scattering processes are observationally allowed. We therefore discuss these only briefly and defer a more detailed study to the future work.

4 RESULTS

4.1 Effective diffusivities and observational constraints

4.1.1 *Effective diffusivities*

Fig. 2 compares the effective diffusivities $\kappa_{\rm eff} \equiv |\boldsymbol{F}|/|\nabla_\parallel e_{\rm cr}|$ from a representative subset of the models in Section 3, at $z = 0$ in a dwarf (**m11i**), intermediate-mass (**m11f**), and MW-mass (**m12i**) galaxy. Among the ET models, as expected, models with larger $f_{\rm turb}$ produce larger $\kappa_{\rm eff}$. Some (e.g. model 'Alfvén-YL02') produce such high $\kappa_{\rm eff} \gg 10^{34}\,{\rm cm^2\,s^{-1}}$ they fall above the plot. Models which ignore anisotropy and/or damping (e.g. 'Iso-K41') produce very low $\kappa_{\rm eff}$; the 'Fast-NoDamp' variant ignoring damping entirely produces $\kappa_{\rm eff} \ll 10^{26}\,{\rm cm^2\,s^{-1}}$, well below the plotted range. In the SC models, $\kappa_{\rm eff}$ is not strongly sensitive to model variations such as the choice of Alfvén speed or equilibrium versus non-equilibrium description, but varies systematically with the strength of turbulent damping (increasing with $f_{\rm cas}$), in an analogous (inverted) manner to the ET models.

There are few other universal systematic trends: (1) $\kappa_{\rm eff}$ tends to rise with galactocentric radius, but the strength of this rise varies widely. (2) There are some radial fluctuations at a given time in $\kappa_{\rm eff}$: there is actually considerably more small-scale scatter than this plot suggests, which depends on how we weight the 'mean' $\kappa_{\rm eff}$, explored below (Fig. 8). (3) In many SC (but not ET) models, the diffusivities are systematically higher in lower mass dwarf galaxies (with lower ρ, $|\boldsymbol{B}|$, $e_{\rm cr}$, etc.).

Some models run are not plotted in Fig. 2, as they simply interpolate between the models shown or give nearly identical results. For example, increasing $\gamma_{\rm L}$ to $\sim$10 in the SC models (Section 3.3.3) simply increases $\kappa_{\rm eff}$ by a factor $\sim$1.5–3 at large radii (and less at $\lesssim$kpc, where ion-neutral damping dominates).

4.1.2 *γ-ray luminosities*

Fig. 3 compares the predicted $\sim$GeV γ-ray emission from each simulation. This was studied in Papers I and II in detail and we follow their methodology, mimicking the compiled (plotted) observations from Lacki et al. (2011), Tang et al. (2014), Griffin et al. (2016), Fu et al. (2017), Wojaczyński & Niedźwiecki (2017), Wang & Fields (2018), Lopez et al. (2018). Briefly, we assume 5/6 of the collisional hadronic losses go to pions, with branching ratio of 1/3 to π^0 that decay to γ-rays with a spectrum giving $\sim$ 70 per cent of the energy at $>$ 1 GeV (Guo & Oh 2008; Chan et al. 2019), and integrate this within apertures ($\sim$ 5–10 kpc) matched to the observations. We similarly compute the central ($\lesssim$2–5 kpc, taken as 1/2 the half-mass radius) projected gas surface density $\Sigma_{\rm central}$, and the luminosity from young/massive stars $L_{\rm SF}$ (using all stars $<$ 100 Myr old, convolved with appropriate stellar population synthesis for their ages and metallicities). The 'calorimetric limit' line denotes the ratio $L_\gamma/L_{\rm sf} = L_{\rm calor}/L_{\rm sf} \sim 2\times10^{-4}$, which corresponds to the assumption that all CR energy injected by SNe is lost collisionally in steady state with a uniform time-constant SFR and SNe rate.

First, let us consider the constant-diffusivity models. These models and variants are the main focus of Papers I and II (with additional simulations and more widely varied assumptions related to streaming and numerics). We echo their conclusion: $\kappa_{29} \sim 3{-}30$ is required to reproduce the observations, with lower $\kappa_{29} \lesssim 1$ producing near-calorimetric predictions even in dwarfs, and $\kappa_{29} \gtrsim 100$ underpredicting L_γ. We also see model $\kappa_{\rm ion\text{-}neutral}$ rather severely overpredicts L_γ, comparable to models with constant $\kappa_{29} \sim 0.5$. We also note (see Papers I and II for further discussion) that adding additional

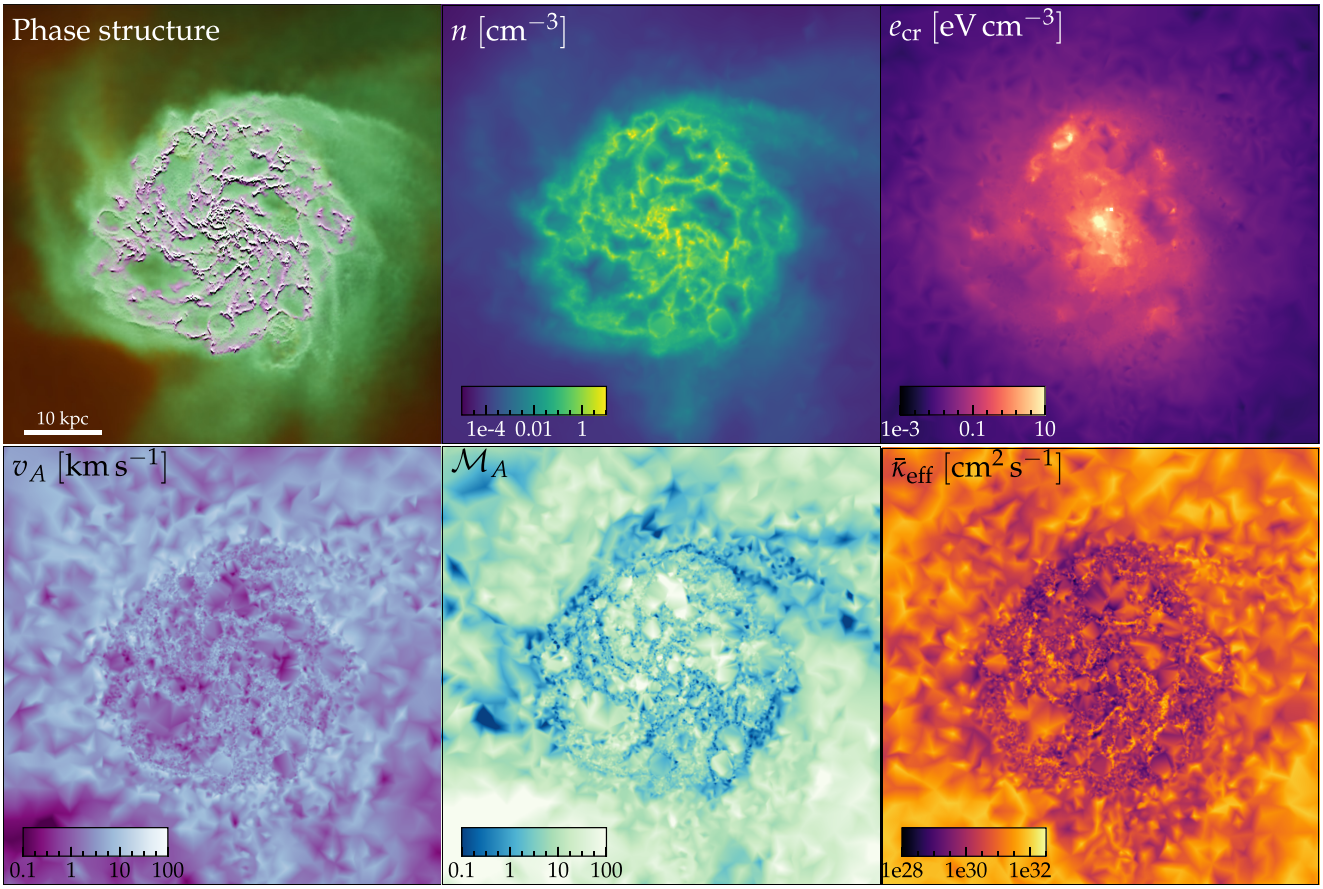

Figure 1. Images of one of our simulated galaxies (**m11f**) at present-day ($z = 0$), in a mid-plane slice with box ~ 60 kpc on a side (see scale-bar), viewed face-on. We show the SC-motived model '$f_{QLT}-100$'. *Top left:* Phase map showing cold neutral (*magenta*, $T \lesssim 8000$ K), warm ionized (*green*; $10^4 \lesssim T \lesssim 10^5$ K), and hot ionized ($T \gtrsim 10^5$ K) gas. *Top centre:* Gas density n. *Top right:* CR energy density e_{cr}. *Bottom left:* Ideal MHD Alfvén speed v_A. *Bottom centre:* Alfvén Mach number $\mathcal{M}_A$. *Bottom right:* Effective diffusivity $\bar{\kappa}_{eff} \equiv |\boldsymbol{F}|/|\nabla_\parallel e_{cr}|$, where $\boldsymbol{F}$ is the local CR flux. Multiphase structure with large fluctuations in turbulent dissipation rates and v_A are evident on scales $\ll$ kpc, while galactic outflows give rise to large $\mathcal{M}_A$ in the CGM and in 'superbubbles' within the disc. These give rise to order-of-magnitude fluctuations in $\bar{\kappa}_{eff}$ on small scales, though $\bar{\kappa}_{eff}$ generally rises outside the galactic disc. The CR energy e_{cr} is smoother, following a radial gradient to first order (as expected), though with a notable 'hotspots' surrounding clustered SNe.

trans-sonic streaming (with $v_{st} \sim v_A^{ideal}$ or $\sim v_A^{ion}$) makes only a small ~ 10 per cent difference to L_γ.

Next, compare ET models: as expected, those with systematically higher κ_{eff} in Fig. 2 produce lower L_γ. Model 'Alfvén-C00' [(i) in Section 3.2] and others with $f_{turb} \gtrsim 100$ in the warm ionized medium (WIM) ($\kappa_{29} \gg 100$) underpredict L_γ: this includes models 'Alfvén-YL02' (ii) and 'Fast-Mod' (iv), which are not shown but fall below the plotted range, and $f_{turb} = 1000$ (vii), which is similar to 'Alfvén-C00' (as expected). Models with $f_{turb} \ll 0.01$, on the other hand, overproduce L_γ, with $\kappa_{29} \lesssim 0.1$ within the galaxy (although κ_{29} varies widely in dwarfs). This includes models 'Iso-K41' (v) and its variants assuming different turbulent spectra or geometries [e.g. models (viii), (ix), (x), (xii), not shown but all similar to 'Iso-K41'], which neglect both the dominant turbulent damping terms and anisotropy of small-scale turbulence in the ISM. For $f_{turb} \sim 0.1-10$, L_γ is broadly similar to observations: this occurs in the ad hoc 'Fast-Max' (iii) and 'Alfvén-Max' ($f_{turb} = 1$; vi) models.

We also see that the 'default' SC model produces excessive L_γ, compared to observations. Varying $v_{st} = v_A^{ion}$ versus v_A^{ideal} has relatively little effect on this conclusion, as does varying the assumed CR energy from $\gamma_L \sim 1-10$ GeV, or adopting non-equilibrium models for κ and v_{st}. Increasing the turbulent damping rate f_{cas} decreases L_γ, with models where $f_{cas} \sim 30$–300 in agreement with the observations. This includes models that increase f_{cas} by a similar factor assuming a different turbulent spectrum (e.g. '$\Gamma_{damp}-$K41').

Fig. 4 also plots L_γ/L_{SF} versus absolute SFR, and L_γ/L_{IR} versus L_{IR}, the total infrared (IR) luminosity (8–1000 μm) computed self-consistently in our simulations by ray-tracing $\sim$100 lines of sight from every star particle (with an input spectrum following the Leitherer et al. 1999 stellar population models for the same age, metallicity, and mass) through the resolved gas and dust in the simulation, assuming an MW-like extinction curve (adopting SMC-like extinction makes little difference) with constant dust-to-metals ratio = 0.4 (see Hopkins et al. 2005). These give somewhat redundant constraints: the same models are (in)consistent with the data in these projections, but they generally show more overlap in the model predictions and are less theoretically well motivated (see Section 5.1.1), so they are less useful for distinguishing models.

4.1.3 *Grammage and residence time*

As discussed in Papers I and II, our comparison to the MW point in Fig. 3 is essentially equivalent to comparing to the observed

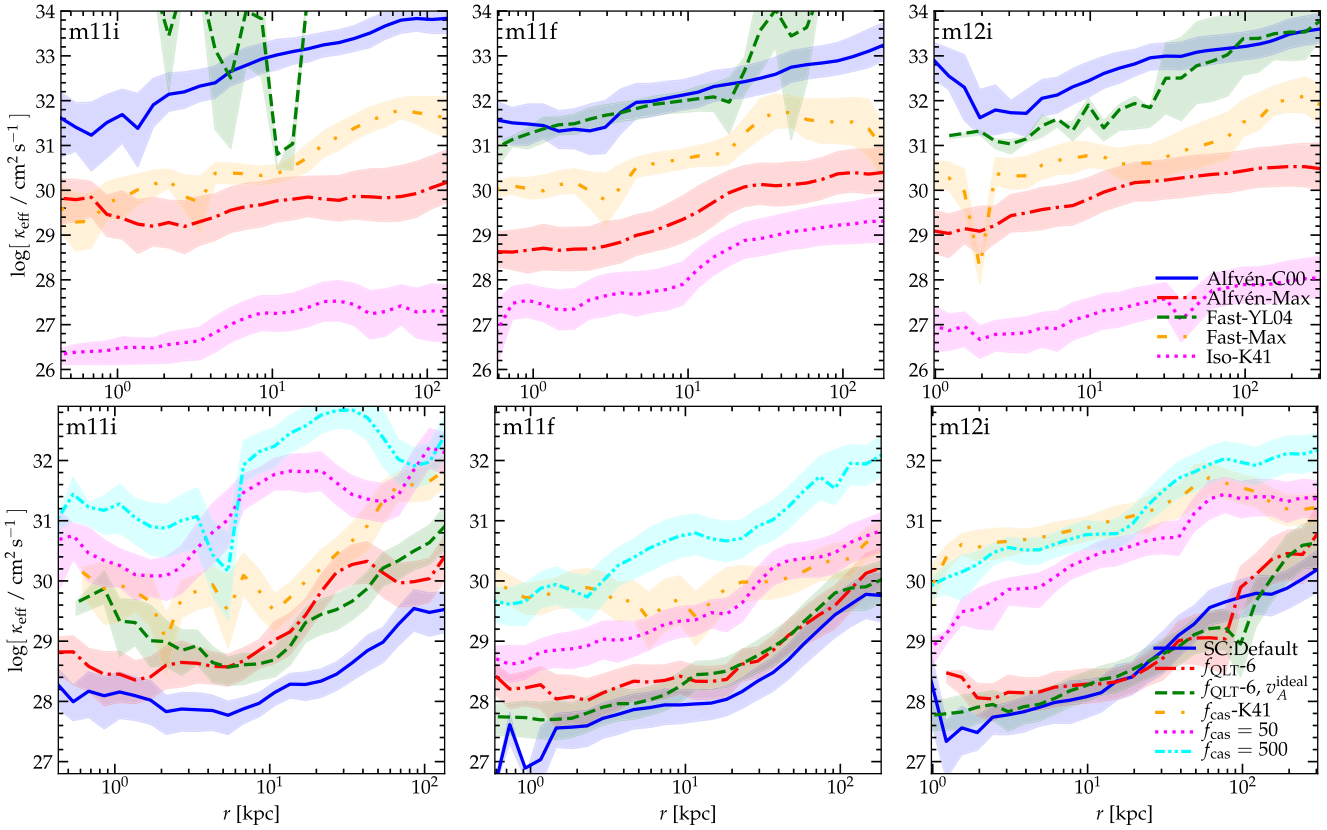

Figure 2. Effective *scattering-weighted* mean parallel CR diffusivity $\kappa_{\rm eff} \equiv |\boldsymbol{F}|/|\nabla_{\parallel} e_{\rm cr}|$ (Section 2.3), as a function of galactocentric radius r, in galaxies **m11i** (dwarf), **m11f** (intermediate/MW/NGC 253-mass), **m12i** (M31-mass) at $z = 0$ (Table 2). We compare some representative models from Table 1 for CR scattering via ET (*top*; Section 3.2) and SC (*bottom*; Section 3.3). Our definition of $\kappa_{\rm eff}$ means this includes both traditional 'diffusion' and 'streaming' terms. Solid lines show the mean $\kappa_{\rm eff}$ in spherical shells at each r, weighted by the contribution of each resolution element to the scattering rate (shaded shows weighted 25–75 per cent range). Diffusivities $\kappa_{\rm eff}$ generally rise with radius r around a given galaxy, or in lower mass dwarf galaxies, as densities ρ and field strengths $|\boldsymbol{B}|$ decrease. Different models considered here produce up to factor $\sim 10^8$ systematic differences in $\kappa_{\rm eff}$ – far larger than any other physical/numerical uncertainties in the models here (see Appendix D). *Top:* Theoretically preferred scattering rates from ET from Alfvén waves ('Alfvén-C00') or fast modes ('Fast-YL04') give large $\kappa_{\rm eff}$: models 'Alfvén-Max' and 'Fast-Max' artificially make the scattering rate much larger ($\kappa_{\rm eff}$ smaller) by neglecting some damping/anisotropy terms, while 'Iso-K41' neglects *all* damping or anisotropy in the turbulence down to $\sim r_L$. *Bottom:* Our 'SC:Default' model (accounting for ion-neutral, turbulent, linear and non-linear Landau damping) produces low $\kappa_{\rm eff}$: multiplying the diffusivity by a factor '$f_{\rm QLT} = 6$' makes little difference owing to non-linear effects (increasing κ produces lower $e_{\rm CR}$, which then re-increases κ in SC models); using the ideal-MHD Alfvén speed $v_A^{\rm ideal}$ instead of the ion speed $v_A^{\rm ion}$ also has weak effects, but $\kappa_{\rm eff}$ can be made larger if $f_{\rm QLT}$ or $f_{\rm cas}$ (turbulent damping rates) are increased by $\sim$100.

grammage in the Galaxy. Specifically, for the MW, quantities like the inferred diffusion coefficient are model dependent: what is most directly constrained by observations like the secondary-to-primary ratios is the effective column density or grammage $X_s \equiv \int_{\rm CR\,path} \rho_{\rm nuclei}\, {\rm d}\ell_{\rm cr} = \int_{\rm CR\,path} \rho_{\rm gas}\, c\, {\rm d}t$ integrated over the path of individual CRs from their source locations to the Earth (with $X_s \sim 5\,{\rm g\,cm^{-2}}$, or $\sim 3\times10^{24}$ nucleons cm^{-2}, measured).[10] If the galaxy is in quasi-steady state with some CR injection rate $\dot{E}_{\rm cr} \propto \dot{E}_{\rm SNe} \propto L_{\rm sf}$ and losses are small ($L_\gamma \ll L_{\rm calor}$), then $e_{\rm cr}(\boldsymbol{x}) \approx \dot{E}_{\rm cr}\,({\rm d}t/{\rm d}^3\boldsymbol{x})$ at some position x (where ${\rm d}t/{\rm d}^3\boldsymbol{x}$ is the residence time of individual CRs in a differential volume element). Using this and the fact that $L_\gamma/L_{\rm calor} = \dot{E}_{\rm coll}/\dot{E}_{\rm cr}$, where $\dot{E}_{\rm coll} = \int {\rm d}^3\boldsymbol{x}\,\Lambda_{\rm coll} = \alpha \int n_{\rm n}\, e_{\rm cr}\, {\rm d}^3\boldsymbol{x}$ (with $\alpha = 5.8\times10^{-16}\,{\rm cm^3\,s^{-1}}$ and $n_{\rm n} = \rho_{\rm nuclei}/m_p$), we obtain

$$X_s^{\infty} \approx 130\,{\rm g\,cm^{-2}}\left(\frac{L_\gamma}{L_{\rm calor}}\right) \qquad (L_\gamma \ll L_{\rm calor}) \tag{10}$$

or $X_s^{\infty} \approx 6\times10^5\,{\rm g\,cm^{-2}}\,(L_\gamma/L_{\rm sf})$ (where X_s^{∞} is the grammage integrated to infinity or 'escape').[11]

We have directly confirmed that this is an excellent approximation in any of our simulations which is remotely consistent with the

[10] Note that the measured grammage we compare to is an energy-weighted average around $\sim$ 1–10 GeV, for which typical estimates in the MW give $\sim$ 2–10 g cm^{-2} (Cowsik, Burch & Madziwa-Nussinov 2014; Korsmeier & Cuoco 2016; Evoli et al. 2017; Amato & Blasi 2018; Kachelrieß & Semikoz 2019).

[11] As $X_s^{\infty} \to \infty$, obviously $L_\gamma/L_{\rm calor} \to 1$, losses become significant, and the linear scaling $X_s^{\infty} \propto L_\gamma/L_{\rm calor}$ in equation (10) breaks down. If we consider a simple slab model we can extend this further, giving

$$X_s^{\infty} \approx 130\,{\rm g\,cm^{-2}}\,\ln\left\{\frac{1}{1-L_\gamma/L_{\rm calor}}\right\}. \tag{11}$$

The simulations do follow this correlation reasonably well for $L_\gamma/L_{\rm calor} \lesssim 1$, but owing to clumpiness (non-'slab' geometric effects) and time variability effects there is no tight correlation once $L_\gamma \gtrsim L_{\rm calor}$. However, these near-calorimetric systems almost always have $X_s \gtrsim 100\,{\rm g\,cm^{-2}}$.

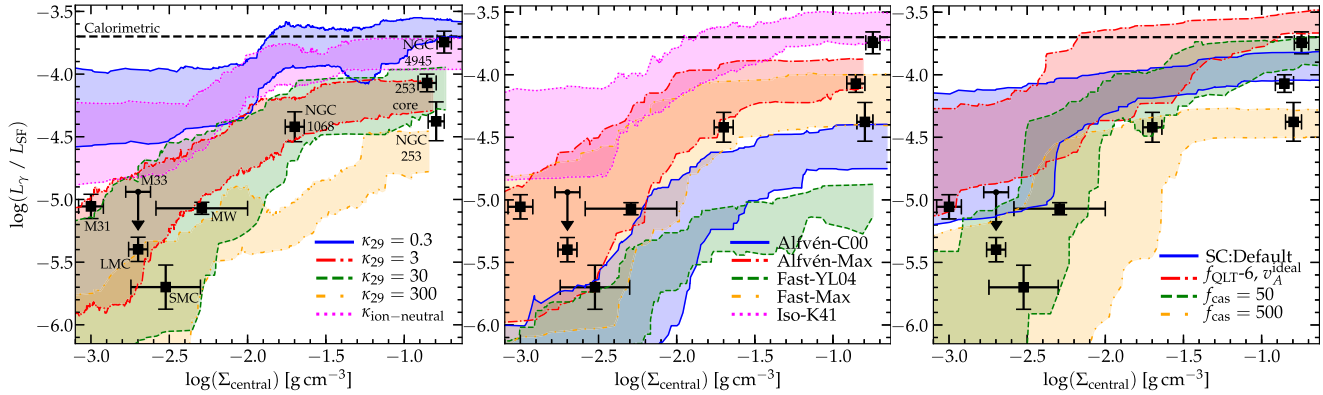

Figure 3. Predicted ratio of γ-ray luminosity from hadronic collisions (L_γ; see Section 4.1.2) to luminosity from star formation/massive stars ($L_{\rm SF}$), as a function of galaxy central gas surface density ($\Sigma_{\rm central}$). Shaded range shows 1σ (~ 68 per cent) inclusion interval of all points measured at uniform time intervals at $z < 1$ (for all **m11i**, **m11f**, **m12i**). Dashed horizontal line is the steady-state calorimetric limit. Black squares compare observations (upper limit is M33). Panel compare subsets of transport models (Table 1). *Left:* Constant-diffusivity (CD; Section 3.1) models. Models with $\kappa_{29} = \kappa_\parallel/10^{29}\ {\rm cm^2\ s^{-1}} \sim 3$–30 agree well with observations. Lower (higher) κ over (under) predicts L_γ. Model '$\kappa_{\rm ion-neutral}$' with $\kappa_{29} = 3$ (0.1) in neutral (ionized) gas only slightly decreases L_γ, relative to models with $\kappa_{29} < 1$ everywhere. *Centre:* ET models. Expected scattering by Alfvénic or fast-mode ET (Alfvén-C00, Fast-YL04) is sub-dominant (underpredicting L_γ), although scattering by fast modes could be important (L_γ similar to observed) under some extreme assumptions (Alfvén-Max, Fast-Max). Model 'Iso-K41' ignores anisotropy and damping of ET, and overpredicts L_γ. *Right:* SC models. 'Default' SC assumptions overpredict L_γ; this is only weakly influenced by the assumed CR energy ($\sim$ 1–10 GeV), choice of Alfvén speed (Section 2.4), and other details. Multiplying the turbulent damping rates by factors $f_{\rm cas} \sim 50$–500, gives good agreement with the observed L_γ.

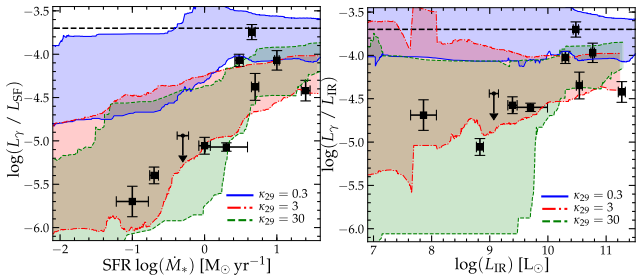

Figure 4. As Fig. 3, comparing $L_\gamma/L_{\rm SF}$ versus the galaxy-integrated SFR $\dot{M}_*$ (*left*) or IR (8–1000 μm) luminosity $L_\gamma/L_{\rm IR}$ versus $L_{\rm IR}$ (*right*; obtained by ray-tracing from each star to a mock observer at infinity assuming an MW-like extinction curve with a constant dust-to-metals ratio equal to the MW value, following Hopkins et al. 2005). Comparing $L_\gamma/L_{\rm SF}$ versus SFR shows essentially identical behaviour to $L_\gamma/L_{\rm SF}$ versus $\Sigma_{\rm central}$ in Fig. 3. Comparing $L_\gamma/L_{\rm IR}$ is less useful: in dwarfs, $L_{\rm IR}/L_{\rm SF}$ declines proportional to the optical/UV attenuation $\tau_{\rm OUV} \approx \kappa_{\rm OUV}\,\Sigma_{\rm central}$, itself proportional to $\Sigma_{\rm central}$, while $L_\gamma/L_{\rm SF}$ similarly scales with $\sim\Sigma_{\rm central}$, so their ratio varies more weakly ($\propto L_{\rm IR}^{0.3}$) and models overlap more heavily. These diagnostics do not rule out any models not already ruled out by the comparison in Fig. 3.

observational constraints, by calculating X_s^∞ following Lagrangian CR trajectories (Fig. 5).[12] To match the constraints at Earth more directly, we have also explicitly calculated $X_s^{(8.1)}$ (or $X_{s,\oplus}$), the grammage from sources to random star particles at the solar circle (8.1 ± 0.1 kpc in the thin disc mid-plane, at $z = 0$) in several of our transport models (for galaxies **m11f** and **m12i**) and in almost all cases find $X_s^{(8.1)} \approx (0.7$–$0.9)\,X_s^\infty$ (since this is well outside the effective radius of star formation in our MW) – a negligible correction compared to other uncertainties here.

We also calculate the *true* 'residence time' $\Delta t_{\rm res}$ of CRs in our simulations by following a random subset of tracer CRs which end up in this mock solar circle at $z = 0$, tracing them back to their time of injection. Note that residence time is only well defined with respect to an observer at a specific location in the galaxy (so we only consider this for our MW-like systems **m11f** and **m12i**), as it diverges for any CRs that escape the galaxy. It also becomes artificially limited by the hadronic loss time-scale $\sim 270\,{\rm Myr}\,(0.1\ {\rm cm^{-3}}/n_{\rm gas})$ when collisional losses become dominant (as $L_\gamma \rightarrow L_{\rm calor}$): indeed, we confirm that all our models with $\Delta t_{\rm res} \gtrsim (1$–$2) \times 10^8$ yr (consistent with loss times for $n \gtrsim 0.1\ {\rm cm^{-3}}$) have $L_\gamma \sim L_{\rm calor}$, and vice versa.[13]

By definition, $\Delta t_{\rm res} = \int_{\rm emission}^{\oplus} {\rm d}t = X_s/(\langle n\rangle\, m_p\, c)$ where $\int_{\rm emission}^{\oplus}$ represents the integral from emission to observation at 'Earth' at $z = 0$, dt is the time along an individual CR trajectory, and $\langle n\rangle \equiv m_p^{-1}\,(\int \rho\,{\rm d}t)/(\int {\rm d}t)$ is a residence-time-weighted average. But in a highly inhomogeneous medium, there is no single $\langle n\rangle$ (and its 'effective' value depends on the transport model). As a result, there is (as one might expect) a broad range of residence times for CRs at the mock observer (with non-trivial 'tails' worth further investigation in future work). Considering just the median at each time, we find that for otherwise 'favoured' models (Alfvén-Max, Fast-Max, $f_{\rm cas}-50$, $f_{\rm QLT}-100$) we obtain median $\Delta t_{\rm res} \sim 3$–50 Myr (and for $f_{\rm cas}-500$, $f_{\rm cas}-$K41 we find $\Delta t_{\rm res} \sim 0.5$–15 Myr) in galaxies **m11f** and **m12i** at times where their $\Sigma_{\rm gas}$ is similar to that of the MW in Fig. 3, matching roughly our expectation given the predicted X_s and a mean $\langle n\rangle \sim 0.1$–1 cm^{-3} typical of the ISM dominating the grammage. But in each of these cases a significant (few per cent or more) fraction of the population seen at the 'observer' has had residence times < 1

[12]Specifically, we re-run the simulation for a short time ~ 300 Myr near $z \approx 0$, with CR tracer particles probabilistically injected every time an SNe injects CR energy (expected number proportional to CR energy injected), each recording its time of injection. Tracers are deleted stochastically with probability equal to the ratio of total catastrophic losses to total CR energy in a cell each time-step, or can 'jump' to neighbour gas cells with probability equal to the fractional CR energy flux from their parent cell to the neighbour (similar to the scheme in Genel et al. 2013).

[13]For example, our 'Iso-K41' and 'SC:Default' models (in **m12i**) give estimated median $\Delta t_{\rm res} \sim 2$–$3 \times 10^8$ yr, but this is primarily limited by hadronic losses in both cases (both have $L_\gamma \sim L_{\rm calor}$). If we ignore the losses for our tracer CRs, we obtain the order-of-magnitude larger $\Delta t_{\rm res} \sim 1$–$4 \times 10^9$ yr.

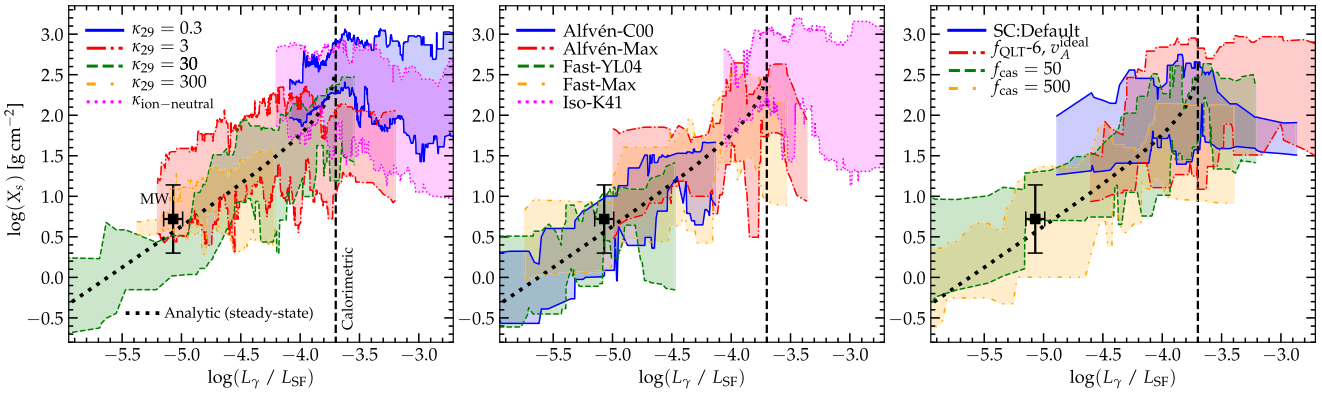

Figure 5. γ-ray luminosity relative to star formation ($L_\gamma/L_{\rm SF}$, models and shaded ranges as Fig. 3) versus CR grammage X_s calculated for an observer far from the galaxy centre, at all simulation times $z < 3$. We label the calorimetric limit and the analytic relation between X_s and $L_\gamma/L_{\rm SF}$ for a homogeneous, steady-state system (equations 10 and 11). Regardless of the CR transport model, the simulations follow $X_s \sim 100\,{\rm g\,cm^{-2}}\,(L_\gamma/L_{\rm SF})$ for $L_\gamma < L_{\rm SF}$, consistent with the MW observations (square labelled). At $L_\gamma > L_{\rm calor}$, X_s saturates (any CRs with higher grammage are lost to collisions before escaping to reach the 'observer'). The scatter is primarily driven by short-time-scale (~ 10 Myr) variations in SFR (i.e. $L_{\rm SF}$) and (to a lesser extent) in L_γ and X_s driven by ISM clumpiness.

or > 50 Myr. All of this is broadly within the range allowed by MW constraints (Strong, Moskalenko & Ptuskin 2007; Putze, Derome & Maurin 2010; Trotta et al. 2011; Aguilar et al. 2016, 2018; Yuan et al. 2017; Kachelrieß & Semikoz 2019). On the other hand (as noted above) the models with $L_\gamma \sim L_{\rm calor}$ all have $\Delta t_{\rm res} \gtrsim 100$ Myr (clearly ruled out), while those with L_γ much less than observed (e.g. 'Alfvén-C00') all have $\Delta t_{\rm res} \lesssim 1$ Myr.

4.1.4 CR energy densities

Fig. 6 compares the radial CR energy density profile averaged in spherical shells,[14] again at $z = 0$, for the same galaxies and models as Fig. 3. For *otherwise fixed galaxy properties*, we expect $e_{\rm cr} \sim \dot{E}_{\rm cr}/(4\pi\, r\, \kappa_{\rm eff}) \propto \kappa_{\rm eff}^{-1}$ in steady state, since the CR flux and hadronic losses must balance the injection by SNe $\dot{E}_{\rm cr}$, on average. In a rough sense, we do see $e_{\rm cr}$ decrease with larger $\kappa_{\rm eff}$ (especially in the constant-κ models), but the trend is weaker and occasionally non-monotonic, owing to the non-linear changes in galaxy properties (e.g. SNe rates) with different κ (see below).

Unlike L_γ, there are no direct observational constraints on $e_{\rm cr}$, except in the Solar neighbourhood (galactocentric $r \sim 8$ kpc) of the MW, where the most current observations indicate $e_{\rm cr} \sim 0.5$–$1.2\,{\rm eV\,cm^{-3}}$ in the diffuse ISM, integrating *all* CRs with energies $\gtrsim 5$ MeV (Webber 1998; Padovani, Galli & Glassgold 2009; Indriolo & McCall 2012; Cummings et al. 2016). This corresponds to $e_{\rm cr} \sim 0.1$–$1\,{\rm eV\,cm^{-3}}$ integrated within a factor of $\sim$10 of 1 GeV. We therefore compare these values to the MW-mass simulations: there are some models which can be ruled out by this constraint, but they are all models *already* ruled out by L_γ or grammage constraints (Fig. 3). Fig. 7 shows this explicitly: we compare more detailed calculations of both $e_{\rm cr}$ and X_s as measured by a mock observer at a random Solar-neighbourhood star, selecting only low-redshift times where the broad galaxy properties (mass and $\Sigma_{\rm central}$ and, as a consequence SFR) are similar to the MW.

For a given CR model, lower mass galaxies exhibit systematically smaller $e_{\rm cr}$ at all radii, as expected given their lower SFRs (hence SNe rates and CR injection rates $\dot{E}_{\rm cr}$), and similar-or-larger $\kappa_{\rm eff}$.

4.1.5 Rigidity dependence of grammage and other properties

It is worth commenting on how the implied grammage and residence time depend on the CR energy $E_{\rm cr} = \gamma_{\rm L}$ GeV or rigidity $\mathcal{R} = \gamma_{\rm L}$ GV. Because our simulations only follow a single bin (so we do not directly evolve high-$\mathcal{R}$ CRs while evolving the $\sim$GeV CRs that dominate $e_{\rm cr}$) we cannot make detailed predictions for this. However, if we assume that higher energy CRs behave as tracers (containing relatively little CR energy) that do not dynamically perturb the galaxies, and neglect losses (valid for $\mathcal{R} \gtrsim 1$ GV), we can predict how $\tilde{\kappa}_{\rm eff}$ and X_s^∞ depend on $\mathcal{R}$ in the different models here.[15] If all else is equal and $\tilde{\kappa}_{\rm eff} = \tilde{\kappa}_{\rm eff}(1\,{\rm GV})\,(\mathcal{R}/{\rm GV})^\delta$ then we simply have $X_s^\infty \propto \mathcal{R}^{-\delta}$. Most analyses of MW observations of, e.g. the B/C ratio, favour $X_s \approx 5\,{\rm g\,cm^{-2}}\,(\mathcal{R}/{\rm GV})^{-(0.5-0.6)}$ (i.e. $\delta \sim 0.5$–0.6) at energies $\sim$1–100 GeV (Ptuskin et al. 2006; Putze et al. 2010; Blasi 2017; Yuan et al. 2017; Aguilar et al. 2018), although systematically varying assumptions about anisotropy, advection/winds, 'halo' size, and source spectral shape can lead to values in the range $\delta \sim 0.3$–0.8 (Maurin, Putze & Derome 2010; Trotta et al. 2011; Blasi 2017).

Although it is commonly assumed that ET models give $\delta = 1/3$ (or $\delta = 1/2$ for a dynamically aligned or Iroshnikov–Kraichnan spectrum), this is only true if anisotropy and damping are totally ignored (as in e.g. our 'Iso-K41' model), which is un-ambiguously ruled out by all other observational constraints. Almost *all* the ET models considered here, give $\delta \lesssim 0$: Alfvén-C00, Alfvén-C00-Vs, Alfvén-Hi, Alfvén-Max all predict $\delta = 0$, while the Alfvén-YL02 model gives negative $\delta = -0.8$. Model Fast-YL04 gives $\kappa_\parallel \propto \mathcal{R}^0$ when collisionless damping dominates and $\propto \mathcal{R}^{-1/6}$ when viscous damping dominates: since viscous damping dominates throughout the ISM and inner CGM, which dominate the residence time, we find, by integrating test particles, an effective $\delta \approx -0.12$ in this model and the related Fast-Mod/Fast-Max/Fast-NoCDamp variations. In short, at energies $\lesssim$TeV (where anisotropy and damping are important),

[14] Because of rapid diffusion, the CR energy density is very similar in cylindrical annuli within the thin disc; see also Fig. 7.

[15] We do this by calculating X_s^∞ for tracer particles (as above) with different $\mathcal{R}$, using the expressions for $\kappa_\parallel(\gamma_{\rm L})$ in the text, then fitting the power-law dependence $X_s^\infty \propto \mathcal{R}^{-\delta}$.

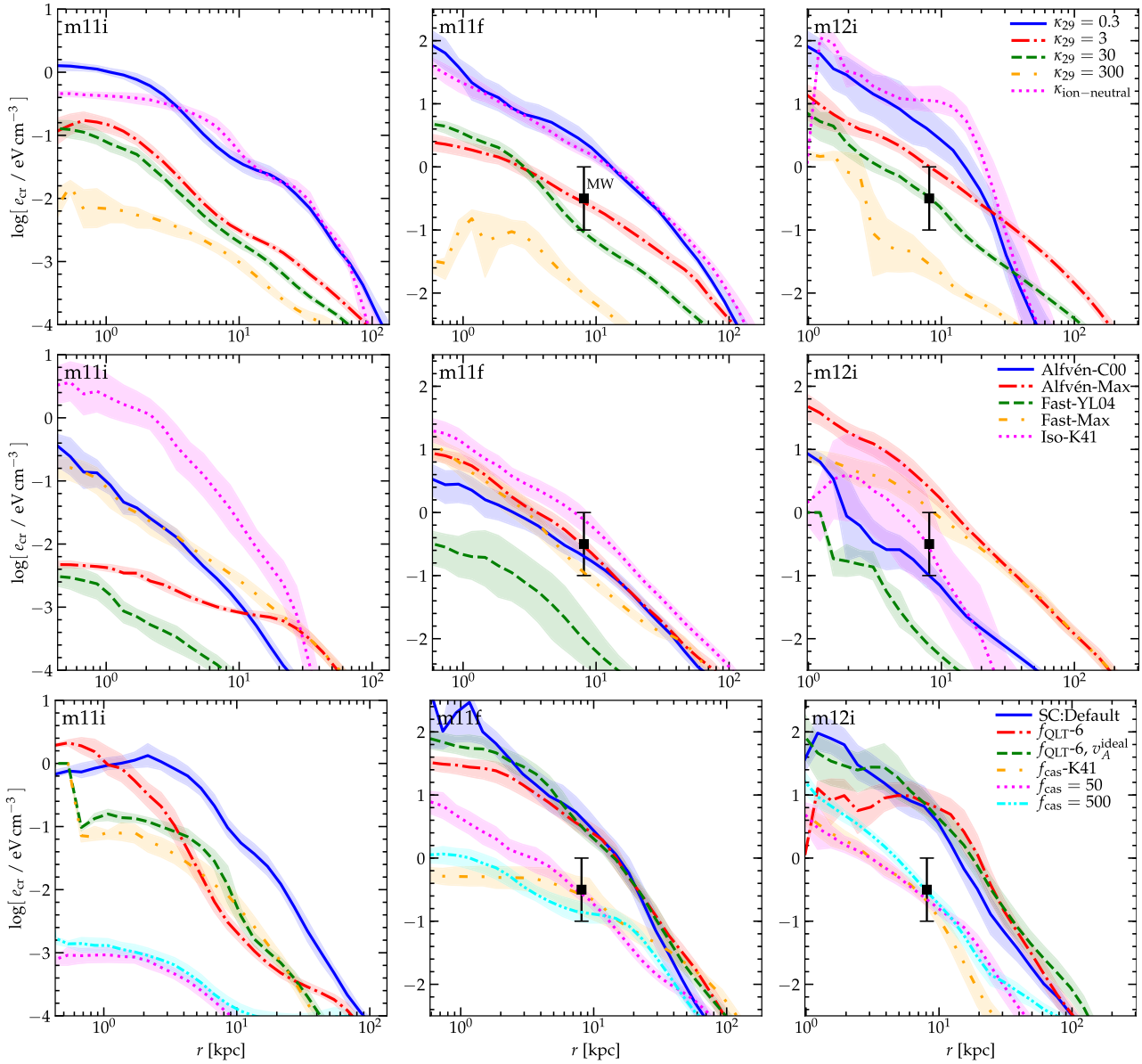

Figure 6. Volume-weighted CR energy density $e_{\rm cr}$ versus galactocentric radius in different transport models (as Fig. 2; see Section 4.1.4). In **m11f** and **m12i**, we note the location and order-of-magnitude observed $e_{\rm cr}$ at the solar circle (error bar). Crudely, $e_{\rm cr}$ decreases as $\kappa_{\rm eff}$ increases in different models. *Top:* CD models. Low (high) $\kappa_{29} \ll 0.3$ ($\gg$30) produce too much (too little) CR confinement and so over (under) predict $e_{\rm cr}$ in MW-like galaxies, consistent with their over (under) prediction of L_γ in Fig. 3. Model $\kappa_{\rm ion\text{-}neutral}$ produces an $e_{\rm cr}$ profile similar to a model with the 'low' ionized gas $\kappa_{29} = 0.1$ everywhere. *Middle:* ET models. Qualitative trends with $\kappa_{\rm eff}$ are similar except model 'Iso-K41' in **m12i** which can produce such efficient CR confinement that CRs lose their energy collisionally, lowering $e_{\rm cr}$. *Bottom:* SC models. These give almost bimodal results in the MW-mass systems, owing to the SC 'runaway' or 'bottleneck' effect where higher $e_{\rm cr}$ produces lower $\kappa_{\rm eff}$ (Section 5.1.3). Transport is 'too slow' in default SC models causing CRs to 'pile up' in excess of observations; $f_{\rm QLT}\,f_{\rm cas} \sim 100$ produces good agreement.

ET models predict the wrong *qualitative* sense of δ, *regardless* of the turbulent spectrum assumed.

On the other hand, in the default SC models here (or those with constant $f_{\rm QLT}$ or $f_{\rm cas}$), $\kappa_\parallel \propto \mathcal{R}^{1/2}$ if turbulent, linear or non-linear Landau damping dominate and $\kappa_\parallel \propto \mathcal{R}^{0-1}$ when ion-neutral damping dominates (0 if $v_A = v_A^{\rm ion}$ dominates over $\kappa_\parallel$, as it often does when ion-neutral damping dominates, 1 otherwise). Since we show below that the grammage and residence times are dominated by the regimes where ion-neutral damping is sub-dominant, we predict an effective $\delta \approx 0.5 \pm 0.1$ for almost all of these models (even models $f_{\rm cas}-$DA and $f_{\rm cas}-$K41, with different turbulent spectra, give $\delta =$ 0.42 and $= 0.36$, respectively).

4.2 Local variations in transport parameters and the 'effective' diffusivity or streaming speed

Having narrowed down the observationally allowed range of ET and SC models, we now explore the distribution of transport parameters in these systems.

4.2.1 Defining 'typical' parameters

Fig. 8 shows $\kappa_{\rm eff}(r)$ and $e_{\rm cr}(r)$, for a representative example of both an ET model ('Fast-Max') and SC model ('SCx100') which produce L_γ and grammage similar to observations (meaning they could, in

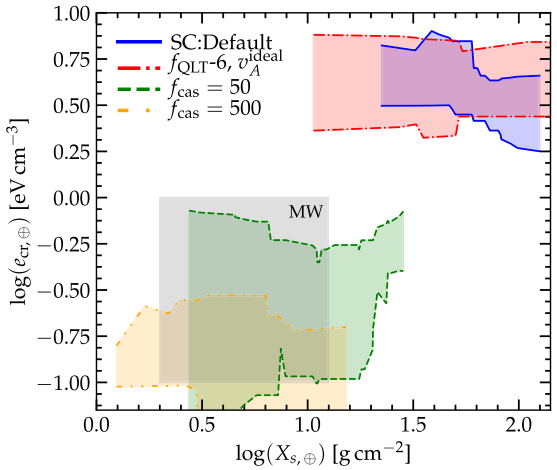

Figure 7. Grammage $X_{s,\odot}$ calculated by following a Monte Carlo subset of CRs from emission to a mock 'observer' at the Solar circle (galactocentric $r = 8.1$ kpc) versus CR energy density in the disc mid-plane at the same location (see Section 4.1.3), sampled over different locations and times at $z < 0.5$ in our **m11f** and **m12i** models selecting times at $z < 1$ where the gas density $\Sigma_{\rm central}$ is similar to the MW value observed (Fig. 3; ~ 0.002–0.01 g cm^{-2}). Shaded grey range shows observationally allowed values for $\sim$GeV CRs. The same models which are consistent with $L_\gamma/L_{\rm SF} \propto X_s$ in Fig. 3 and $e_{\rm cr}$ in Fig. 6 are consistent with the grammage/residence time constraints, for galaxies at times similar to the MW. We show a subset of SC models but have considered additional ET and CD models and reach the same conclusion.

principle, represent the dominant CR scattering). We determine the median and scatter in each annulus with various different weights, e.g. weighting each cell by the local gas mass ($\rho\,{\rm d}^3\boldsymbol{x}$), volume (${\rm d}^3\boldsymbol{x}$), CR energy ($e_{\rm cr}\,{\rm d}^3\boldsymbol{x}$), grammage or contribution to L_γ ($\propto e_{\rm cr}\,\rho_{\rm gas}\,{\rm d}^3\boldsymbol{x}$), CR scattering rate ($\propto (e_{\rm cr}/\kappa)\,{\rm d}^3\boldsymbol{x}$), or CR residence time ($\propto (e_{\rm cr}\,{\rm d}^3\boldsymbol{x})\,(e_{\rm cr}\,{\rm d}r/|\boldsymbol{F}|)$). Fig. 1 highlights local variations in $e_{\rm cr}$ and $\kappa_{\rm eff}$ by showing a 2D map of their local values, in a slice through the galaxy.

Within the galaxy, we see the resulting 'typical' $\kappa_{\rm eff}$ differs by as much as ~ 2 dex (in the CGM, the differences are ~ 0.5–1 dex). This owes to inhomogeneity in the plasma properties inside the ISM, discussed below (Section 5.1.3) and which, in these CR transport models, directly translates to large (order-of-magnitude) local variations in $\kappa_{\rm eff}$ and $v_{\rm st}$. Weighting by, e.g. volume, favours diffuse ISM. Weighting by scattering rates or residence times, $\propto 1/\kappa_{\rm eff}$, selects the *lowest* local values of $\kappa_{\rm eff}$, as relevant to the 'residence' or 'escape' time in an inhomogeneous medium, which is dominated by the regions with the slowest CR propagation. Fundamentally, different 'weights' correspond to different questions: observational constraints on L_γ and grammage are sensitive to residence-time-weighted transport parameters, while the median CR energy density and effects of CRs on pressure support of the CGM and ISM are sensitive to the ISM mass and volume-weighted parameters.

We also see this inhomogeneity reflected in significant time-variation in Fig. 9, even averaging within annuli. Relatively large-scale structure in $\kappa_{\rm eff}$ at a given radius (dominated by spiral arms or large cloud complexes or superbubbles) can still be somewhat transient, producing factor $\sim$3–10 changes in the mean $\kappa_{\rm eff}$ within an annulus over a galactic dynamical time (while smaller structures vary on smaller time-scales). Galactic-scale 'events' (a burst of star formation and associated outflow) can produce large coherent changes in $e_{\rm cr}$ and $\kappa_{\rm eff}$.

This explains much of why there is not a trivial one-to-one linear relation between $\kappa_{\rm eff}$ and L_γ in Figs 2 and 3, in the SC and ET models.

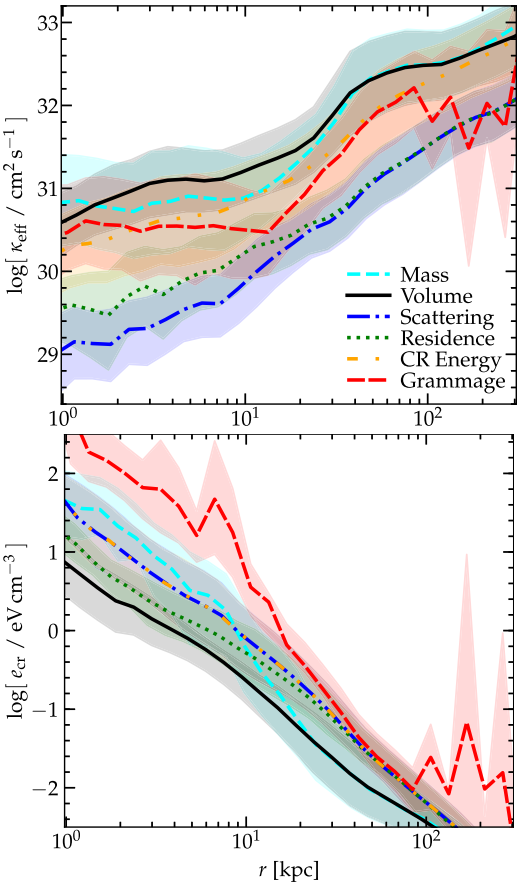

Figure 8. Radial profile of $\kappa_{\rm eff}$ (*top*; as Fig. 2) and $e_{\rm cr}$ (*bottom*; as Fig. 6), in one example consistent with observations (**m12i** in SC model '$f_{\rm QLT}-100$'). We calculate the profiles weighting each resolution element by different quantities in each radial annulus (Section 4.2.1, averaged over all times $z < 0.5$): gas mass, volume, CR scattering rate, CR residence time, CR energy, grammage (or equivalently contribution to L_γ). *Top:* The 'mean' $\kappa_{\rm eff}$ (at a fixed radius and time) can vary systematically by factors up to $\sim$100 based on weight, owing to the very large local variations in the ISM/CGM (Fig. 1). Weighting by scattering rate or residence time ($\propto 1/\kappa$) biases towards the lowest κ regions, where CRs can be 'trapped', while volume-weighting gives the highest κ and others lie in-between. Differences are smaller in the CGM (where e.g. density differences in phases are less extreme), but still factor $\sim$10. *Bottom:* Because of rapid diffusion, differences in $e_{\rm cr}$ are smaller (it is smoother; see Fig. 1), but still significant, as weighting by e.g. total grammage ($\propto e_{\rm cr}\,\rho\,{\rm d}^3\mathbf{x}$) biases to the densest gas with the highest $e_{\rm cr}$.

Some of these models can produce very large volume or L_γ-weighted $\kappa_{\rm eff}$, but in the central few kpc of the galaxy (which dominate L_γ) the residence-time or scattering-rate weighted $\kappa_{\rm eff}$ is much lower, producing larger L_γ. Some of this variation also translates to $e_{\rm cr}$, although the diffusive nature of CR transport reduces the variations here.

4.2.2 *Diffusion versus streaming*

Fig. 10 compares $\kappa_{\rm eff}(r)$ with different weights like Fig. 8, but extends this to dwarf and intermediate-mass galaxies, and also compares the effective streaming speed $\bar{v}_{\rm st,\,eff}(r)$. Recall (Section 2.3) we can freely translate locally between the two using $\bar{v}_{\rm st,\,eff} \equiv \kappa_{\rm eff}/(\gamma_{\rm cr}\,\ell_{\rm cr})$. Fig. 10 considers $\bar{v}_{\rm st,\,eff}$ in absolute units as well as relative to $v_A^{\rm ideal}$ and $v_A^{\rm ion}$.

First, we see that the local and systematic variations (weight-dependence) in $\kappa_{\rm eff}$ *within a single galaxy* discussed above extend

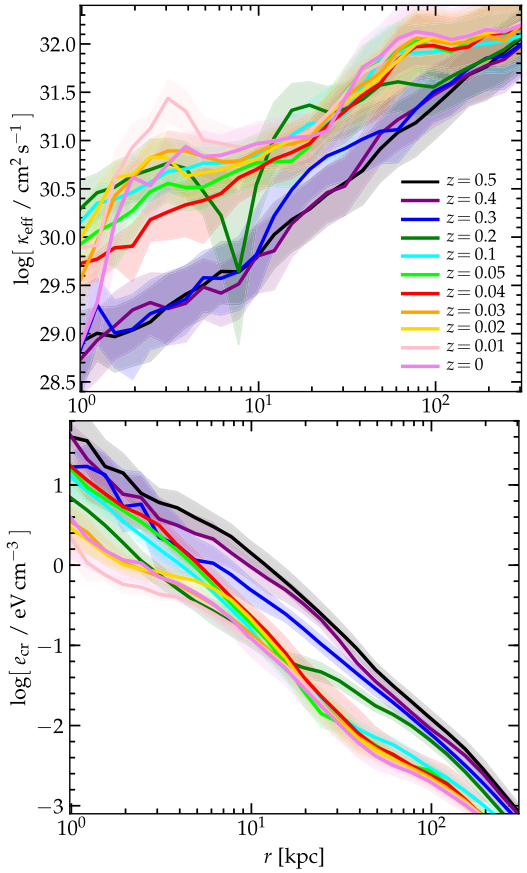

Figure 9. Time-dependence of $\kappa_{\rm eff}$ (scattering-rate-weighted) and $e_{\rm cr}$ (volume-weighted). We plot profiles of both in **m12i** SC model '$f_{\rm QLT}-100$' as Fig. 8, but sampling different times at $z < 0.5$ (different colours; note the time/redshift spacing is not uniform). There is considerable variation in time, which is not simply a continuous systematic evolution but reflects substantial changes over time as bar and spiral arms, phase structure and presence/absence of superbubbles, and periods of elevated star formation (e.g. associated with higher $e_{\rm cr}$ at $z\sim0.3$–0.5) and galactic outflow appear and recede.

to all galaxies simulated. They also do not vanish or significantly decrease if we consider $\bar{v}_{\rm st,\,eff}$ or $\bar{v}_{\rm st,\,eff}/v_A$ instead of $\kappa_{\rm eff}$. Likewise, systematic galaxy-to-galaxy variations in $\kappa_{\rm eff}$ (being larger in dwarfs) appear in $\bar{v}_{\rm st,\,eff}$ as well. In other words, these results are not simply an artefact of parametrizing the transport with $\kappa_{\rm eff}$ instead of $\bar{v}_{\rm st,\,eff}$.

Secondly, we see that, for a given model and weight (usually), $\kappa_{\rm eff}$ is approximately independent of r within the galaxy (within a few kpc), but then rises at larger r (in the CGM), while $\bar{v}_{\rm st,\,eff}$ depends on r within the galaxy but is less-strongly r-dependent in the CGM.

Thirdly, we see that $\bar{v}_{\rm st,\,eff}$ in absolute units is actually closer to r-independent (and exhibits weaker systematic weight-dependence), compared to $\bar{v}_{\rm st,\,eff}/v_A^{\rm ideal}$ or $\bar{v}_{\rm st,\,eff}/v_A^{\rm ion}$, even though the SC simulations plotted assume $v_{\rm st} = v_A^{\rm ion}$. In other words, because $\kappa_\parallel$ is non-zero, we have $\bar{v}_{\rm st,\,eff} \approx v_{\rm st} + \kappa_\parallel/(\gamma_{\rm cr}\,\ell_{\rm cr}) \neq v_{\rm st}$.

Finally, we stress that even if the average $\kappa_{\rm eff}$ or $v_{\rm st,\,eff}$ were approximately constant across galactocentric radius and time, the transport equations being integrated (especially for SC models) do not actually have the same form as a 'true' diffusion or streaming/advection equation (see Appendix B3). Thus, while $\kappa_{\rm eff}$ or $v_{\rm st,\,eff}$ are useful parameters and can guide our intuition regarding transport timescales, equilibrium fluxes, etc., care is required in their interpretation.

4.3 Redshift dependence and effects on galaxy evolution

In future work, we will explore in detail the effect of different CR models on *galaxy* properties, e.g. how they influence galactic star formation and ISM/CGM properties. Because our focus in this work is the observational constraints on CR transport models, we only briefly discuss galaxy properties here in so far as it can provide additional constraints. In Papers I and II, we showed using 'constant-diffusivity' models that entirely turning on/off CRs, or changing κ by factors of $\sim$1000, makes only a modest (albeit non-negligible and potentially important) difference to global galaxy properties. We found that the strongest effects due to CRs (choosing the 'most optimal' diffusivity) occur around MW-mass at $z\sim0$, and even there it typically results in factor $\lesssim$2–3 differences in e.g. galaxy stellar masses. This is not sufficiently large to obviously rule out a specific CR transport model or diffusivity (because, e.g. changing the mean mechanical energy per SNe by a similar factor, easily allowed by observations, would result in a similar effect). Among the models studied here which are allowed by γ-ray observations, we generally find effects on galaxy formation 'in between' the 'no CR' and 'largest CR effects' models from Paper II. We also find (consistent with Paper II) that effects of CRs on galaxy properties are weaker at high redshifts (in every model considered here), owing to relatively higher ISM/CGM pressures. We therefore conclude that the indirect effects of CRs on bulk galaxy properties do not strongly constrain the CR transport models of interest.

5 DISCUSSION

5.1 The need for 'fast' transport and cosmological simulations with resolved ISM phases

5.1.1 Favoured transport parameters: an analytic toy model

The total (galaxy-integrated) CR collisional loss rate is $\dot{E}_{\rm coll} \equiv \int {\rm d}^3\boldsymbol{x}\,\Lambda_{\rm coll}(n_{\rm gas},\,e_{\rm cr})$. In Paper II, we developed a simple toy model for a constant isotropically averaged diffusivity $\tilde{\kappa}_{\rm eff} \sim \kappa_{\rm eff}/3 \equiv \tilde{\kappa}_{29}\,10^{29}\,{\rm cm^2\,s^{-1}}$ (or $\tilde{v}_{\rm st,\,eff} \sim \bar{v}_{\rm st,\,eff}/3 \equiv \tilde{v}^{\rm st}_{1000}\,1000\,{\rm km\,s^{-1}}$) in a disc+halo system, with a steady-state star formation and SNe rate, hence constant $\dot{E}_{\rm cr} \approx 0.1\,\epsilon_{\rm SNe}\,\dot{M}_*$ (where $\epsilon_{\rm SNe}\sim10^{51}\,{\rm erg}/100\,{\rm M_\odot}$ is the energy per unit stellar mass in SNe). If the CRs are confined (not free-escaping), diffusion is relatively fast (compared to e.g. bulk gas motion), the SFR (hence CR injection) is centrally concentrated compared to the size of the CR halo, and collisional losses are small, then in steady state the CR energy density should scale as $e_{\rm cr} \sim \dot{E}_{\rm cr}/(4\pi\,\tilde{\kappa}_{\rm eff}\,r) \sim \dot{E}_{\rm cr}/(4\pi\,\tilde{v}_{\rm st}\,r^2)$. If the disc+halo follows a realistic extended profile with most of the gas mass $M_{\rm gas}$ in a half-mass radius $\ell_{\rm gas}$ and central surface density $\Sigma_{\rm gas}$, then (performing the integrals exactly for a thin, exponential disc in a power-law halo following Paper II):

$$\frac{\dot{E}_{\rm coll}}{\dot{E}_{\rm cr}} \approx \frac{L_\gamma}{L_{\rm calor}} \sim \frac{0.15}{\tilde{\kappa}_{29}}\left(\frac{\Sigma_{\rm gas}\,\ell_{\rm gas}}{0.01\,{\rm g\,cm^{-2}\,kpc}}\right) \sim \frac{0.06}{\tilde{v}^{\rm st}_{1000}}\left(\frac{\Sigma_{\rm gas}}{0.01\,{\rm g\,cm^{-2}}}\right), \tag{12}$$

or equivalently (using $L_{\rm calor} \approx 2\times10^{-4}\,L_{\rm sf}$)

$$\frac{L_\gamma}{L_{\rm sf}} \sim \frac{3\times10^{-5}}{\tilde{\kappa}_{29}}\left(\frac{\Sigma_{\rm gas}\,\ell_{\rm gas}}{0.01\,{\rm g\,cm^{-2}\,kpc}}\right) \sim \frac{10^{-5}}{\tilde{v}^{\rm st}_{1000}}\left(\frac{\Sigma_{\rm gas}}{0.01\,{\rm g\,cm^{-2}}}\right). \tag{13}$$

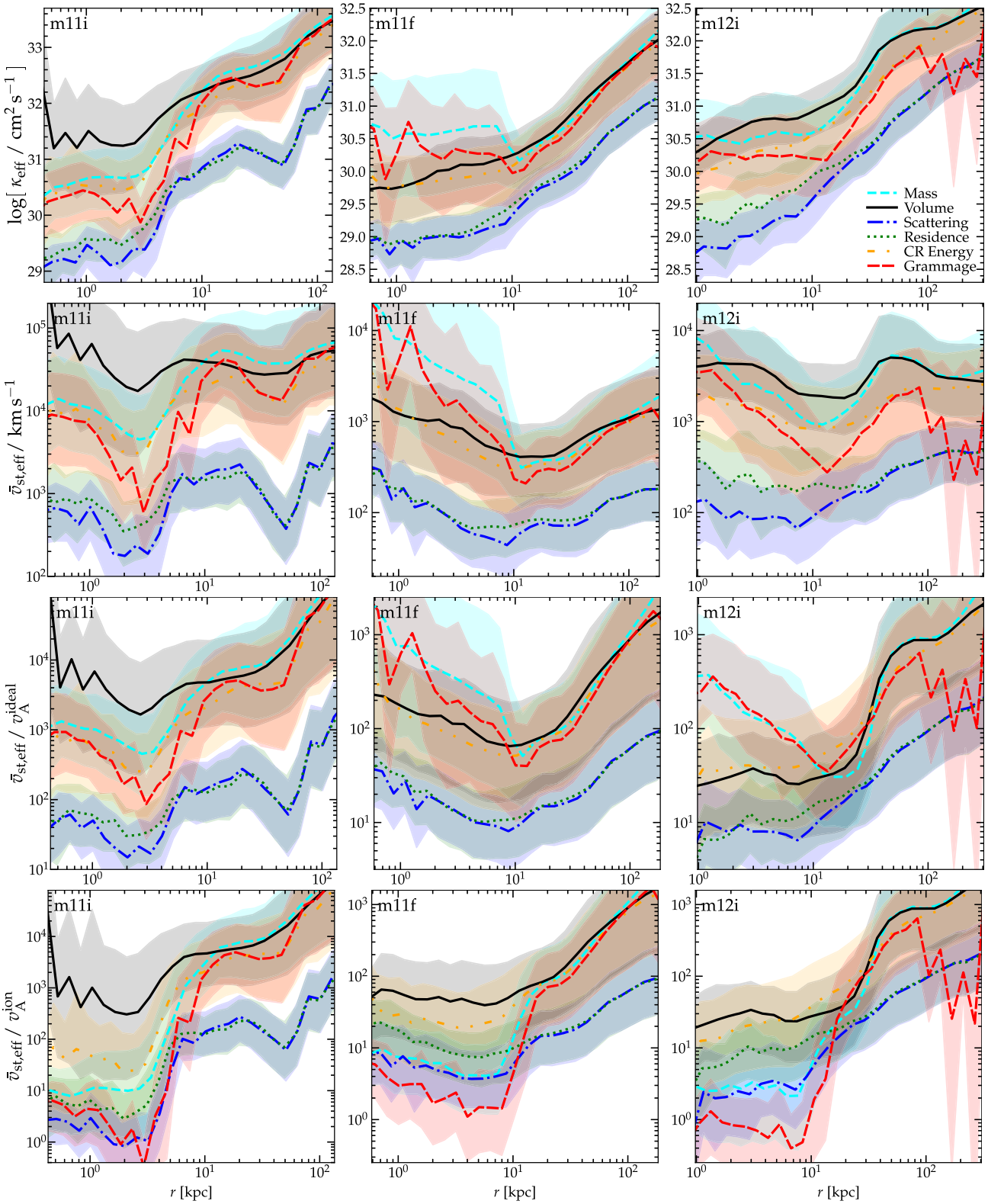

Figure 10. Radial profile of transport parameters in SC model '$f_{\rm QLT}-100$' in different galaxies versus weight (as Fig. 8), see Section 4.2.2. *Top:* Effective 'diffusivity' $\kappa_{\rm eff} \equiv |\boldsymbol{F}|/|\nabla_{\parallel} e_{\rm cr}|$. *Second:* Effective 'streaming speed' $\bar{v}_{\rm st,\,eff} \equiv |\boldsymbol{F}|/h_{\rm cr}$. *Third:* Effective streaming speed in units of local ideal-MHD Alfvén-speed $\bar{v}_{\rm st,\,eff}/v_A^{\rm ideal}$. *Bottom:* Effective streaming speed in units of local ion Alfvén-speed $\bar{v}_{\rm st,\,eff}/v_A^{\rm ion}$ ($v_A^{\rm ion} \approx f_{\rm ion}^{-1/2}\, v_A^{\rm ideal}$). In all cases, the choice of weight has similar (large) effects: this reflects genuine inhomogeneity, not the particular diagnostic. Diffusivity $\kappa_{\rm eff}$ is reasonably constant within a single galaxy ($r \lesssim$ a few kpc) but rises with r in the CGM (by factors $\sim$100–1000 at the virial radius); the scattering-weighted $\kappa_{\rm eff}$ also depends surprisingly weakly on which galaxy we consider. The absolute $\bar{v}_{\rm st,\,eff}$ is much closer to r-independent, though the scattering-rate-weighted value $\sim$ 100–1000 km s^{-1} depends more strongly systematically on galaxy type. Considering $\bar{v}_{\rm st,\,eff}$ in units of $v_A^{\rm ideal}$ or $v_A^{\rm ion}$ increases the scatter/radius dependence/systematic variations between galaxies: it is not accurate to simply think of 'super-Alfvénic streaming' arising from SC as some multiple of v_A.

In terms of the grammage X_s^{∞}, this gives

$$\frac{X_s^{\infty}}{\mathrm{g\,cm^{-2}}} \sim \frac{20}{\tilde{\kappa}_{29}} \left(\frac{\Sigma_{\rm gas}\,\ell_{\rm gas}}{0.01\,\mathrm{g\,cm^{-2}\,kpc}} \right) \sim \frac{6}{\tilde{v}_{1000}^{\rm st}} \left(\frac{\Sigma_{\rm gas}}{0.01\,\mathrm{g\,cm^{-2}}} \right). \quad (14)$$

The assumption that losses are small means this applies when $\dot{E}_{\rm loss}/\dot{E}_{\rm coll} \ll 1$; losses will saturate at the calorimetric limit $\dot{E}_{\rm loss} \approx \dot{E}_{\rm coll}$. This simple estimate gives a surprisingly good estimate of the full simulation prediction for $L_\gamma/L_{\rm sf}$ for our constant-κ models (assuming $\tilde{\kappa}_{\rm eff} \sim \kappa_\parallel/3$) in Fig. 3.

Moreover, if we assume we are in an MW-like galaxy, with a 'solar circle' at $r_{\rm obs} \approx 8$ kpc, we can also estimate the median CR energy density and CR residence time[16] seen by a mock observer:

$$\left.\frac{e_{\rm cr}}{\mathrm{eV\,cm^{-3}}}\right|_{\odot} \sim \frac{2}{\tilde{\kappa}_{29}} \left(\frac{R_{\rm SNe,\,MW}}{1/30\,\mathrm{yr}} \right), \quad (15)$$

$$\left.\frac{\Delta t_{\rm res}}{\mathrm{Myr}}\right|_{\odot} \sim \frac{25}{\tilde{\kappa}_{29}} \left(\frac{r_{\rm obs}^2 - r_{1/2}^2}{(8\,\mathrm{kpc})^2 - (5\,\mathrm{kpc})^2} \right), \quad (16)$$

where $R_{\rm SNe,\,MW}$ is the MW (Galaxy-integrated) SNe rate $\sim 1/30\,\mathrm{yr}$.

Noting that the MW has an observed central $\Sigma_{\rm gas} \sim 20\,\mathrm{M_\odot\,pc^{-2}} \sim 0.004\,\mathrm{g\,cm^{-2}}$ and $\ell_{\rm gas} \sim 5$ kpc, reproducing the observed MW grammage $X_s \sim 3$–$10\,\mathrm{g\,cm^{-2}}$, $L_\gamma/L_{\rm sf} \sim 0.03$, $e_{\rm cr} \sim 0.1$–$1\,\mathrm{eV\,cm^{-3}}$, or $\Delta t_{\rm res} \sim 5$–$20\,\mathrm{Myr}$ all require $\tilde{\kappa}_{29} \sim$ a few. This is the median of our 'favoured' values in Table 1.

This also neatly illustrates the degeneracy between inferred diffusivity and 'halo size' in simpler leaky-box models: if the CRs escape at some height $h < \ell_{\rm gas}$ (truncating the integral above), it is roughly equivalent to replacing $\ell_{\rm gas} \rightarrow h$ in the calculation above, and for a fixed $L_\gamma/L_{\rm sf}$ or X_s, we have an inferred $\kappa \propto h$. As soon as we abandon the assumption of a 'leaky box' or 'flat halo' with $h < 1$ kpc, all of the observations require similar, relatively 'fast' transport speeds.

5.1.2 *Scalings of gamma-ray luminosity with galaxy properties*

The simple model in Section 5.1.1 and equation (13) naturally explains the trend of $L_\gamma/L_{\rm SF} \propto \Sigma_{\rm gas}$ at low $\Sigma_{\rm gas}$ seen in Fig. 3, as $L_\gamma \propto X_s \propto \Sigma_{\rm gas}$ – i.e. for a similar transport speed, the grammage X_s (and therefore L_γ produced by collisions) simply scales with the galactic column density.

In contrast, the trend of $L_\gamma/L_{\rm SF}$ with $L_{\rm SF}$ or $\dot{M}_*$ in Fig. 4 is closer to $L_\gamma/L_{\rm SF} \propto \dot{M}_*^{0.7}$. This follows from global galaxy scalings like the Schmidt–Kennicutt relation $\dot{\Sigma}_* \propto \Sigma_{\rm gas}^{1.4}$ seen in both nature and these simulations (Kennicutt 1998; Orr et al. 2018), which (with equation 13) gives $L_\gamma/L_{\rm SF} \propto \dot{M}_*^{0.7}/\tilde{\kappa}_{29}$.

If we assume steady state with a constant SFR, then the total IR luminosity is determined by the fraction of optical/UV light absorbed and re-emitted: $L_{\rm IR}/L_{\rm SF} \approx (1 - \exp\,[-\kappa_{\rm OUV}\,\Sigma_{\rm gas}])$ where $\kappa_{\rm OUV} \sim 1000\,\mathrm{cm^2\,g^{-1}}\,(Z/Z_\odot)$ is the flux-averaged optical/UV opacity (scaling with galaxy metallicity Z). In dwarfs and the MW where $L_{\rm IR} \lesssim L_{\rm SF}$ this gives $L_{\rm IR}/L_{\rm SF} \sim \kappa_{\rm OUV}\,\Sigma_{\rm gas}$. Combining with equation (13), we have $L_\gamma/L_{\rm IR} \sim 3 \times 10^{-5}\,\tilde{\kappa}_{29}^{-1}\,(\ell_{\rm gas}/10\,\mathrm{kpc})\,(Z_\odot/Z)$, which is very weakly dependent on galaxy properties (both $\ell_{\rm gas}$ and Z scale $\propto M_*^{0.2-0.3}$, and their scalings cancel here; see Kewley & Ellison 2008; Hall et al. 2012). In short, the fact that $L_\gamma/L_{\rm IR}$, while clearly not constant, depends only weakly on $L_{\rm IR}^{0.2-0.3}$ (Fig. 4) – i.e. that the $L_\gamma - L_{\rm IR}$ relation is closer to linear than the L_γ–SFR relation, trivially follows from the fact that *both* the grammage X_s (which is proportional to L_γ) and OUV optical depth τ (proportional to $L_{\rm IR}$) scale with $\Sigma_{\rm gas}$.

Again, reproducing any of the observed trends requires similar $\tilde{\kappa}_{29} \sim$ a few.

5.1.3 *Importance of cosmological simulations and resolved ISM/CGM phases*

Although the simple analytic scalings above can explain many qualitative phenomena, we also identify in our simulations a number of important effects which can only be properly captured in cosmological simulations with *resolved* ISM phases. These include the following:

(i) Extended haloes: Galaxies have extended gaseous haloes reaching to > 100 kpc, containing most of the gas mass in relatively slowly falling power-law density profiles (e.g. isothermal $\rho \propto r^{-2}$, as opposed to exponential). In *every* physically plausible model we consider, the $\sim$ GeV CRs remain confined/coupled in the halo out to $\gtrsim R_{\rm vir}$ (mean free paths are $\lambda_{\rm mfp} \sim 3\,\kappa/c \sim 0.003\,\kappa_{29}$ kpc, compared to ~ 100 kpc halo scale-lengths). The galaxy and even 'inner' CGM halo at $\lesssim 10$ kpc is not a 'leaky box' or 'flat halo' with simple escape outside some volume.

(ii) Clumpiness: At high $\tilde{\kappa}_{\rm eff}$, ISM 'clumping' does not strongly alter L_γ because CRs rapidly move through dense gas. But if $\tilde{\kappa}_{\rm eff} \lesssim 10^{27}\,\mathrm{cm^2\,s^{-1}}$ locally, then CR diffusion/escape times ($\sim \ell^2/\kappa$) becomes shorter than (a) the dynamical times ($\sim 1/\sqrt{G\,\rho}$) of large ($\gtrsim 100\,\mathrm{pc}$) GMC complexes and (b) CR collisional loss times ($\sim 40\,n_1^{-1}$ Myr). Thus, CRs get 'captured' in dense clumps, producing order-of-magnitude higher L_γ.

(iii) Multiphase neutral gas: If the neutral gas is *bounded* (e.g. in clouds or a thin disc) by ionized gas, then even if $\tilde{\kappa}_{\rm eff} \rightarrow \infty$ in that neutral gas, the CR energy density $e_{\rm cr}$ becomes locally constant at a value $\langle e_{\rm cr} \rangle$ determined by the 'boundary condition' value of $e_{\rm cr}$ in the ionized medium. If $\tilde{\kappa}_{\rm eff}$ is low in the ionized gas, the CRs are therefore 'trapped' regardless of $\tilde{\kappa}_{\rm eff}$ in the cold/neutral phase. Thus, the total residence time in dense gas can be large, in principle, even if the local diffusivity in said gas is also large.

(iv) Halo 'collapse': As shown in Paper II, if CRs efficiently escape the disc to $\gtrsim 10$ kpc in intermediate and MW-mass systems, they provide substantial pressure support to the halo gas, which in turn suppresses accretion leading to significantly less dense gas *in the disc* at $z \approx 0$, which suppresses L_γ further. But if they *cannot* escape to $\gtrsim 10$ kpc, the halo 'collapses' and produces more efficient cooling and denser gaseous discs in MW-mass systems, non-linearly raising L_γ.

(v) Self-confinement 'runaway' or 'bottleneck': In SC models, the diffusivity/streaming speed scales inversely with $e_{\rm cr}$ (i.e. the absolute CR *flux* is bottlenecked by the self-excited waves). Thus, if $e_{\rm cr}$ builds up to large ISM values even briefly, the effect rapidly runs away, as it restricts its own transport. A number of other non-linear effects can further exacerbate this: for example, if $P_{\rm cr}$ begins to dominate pressure support in the WIM or inner CGM, then turbulence is generally weaker (as CRs suppress rapid gas cooling/collapse and star formation), hence $\Gamma_{\rm turb}$ and κ become smaller still. These produce large local fluctuations in diffusivity/streaming speed.

(vi) Clustered SNe: In a resolved ISM, SNe are strongly clustered in space and time and associated with denser, star-forming regions.

[16]For residence time, we model CR injection as a Gaussian with initial half-mass radius $r_{1/2} = 5$ kpc, motivated by the stellar (and SNe Ia) scale length in the MW (adopting the scale length for young-stars, for core-collapse, gives $r_{1/2} \approx 3$ kpc), diffusing isotropically, then calculate the median time-since-injection of all CRs in a shell $r_{\rm obs} \approx 8$ kpc in steady state.

This enhances L_γ directly, but more importantly leads to *locally* large $e_{\rm cr}$ which can trigger the SC runaway discussed above.

(vii) Tangled fields: Magnetic fields are highly 'tangled' (Su et al. 2018; Ji et al. 2020), reducing $\tilde{\kappa}_{\rm eff}$. And in some cases (e.g. strong oblique shocks), perpendicular $\boldsymbol{B}$-fields enhance CR 'trapping' in high-density gas, which can enhance L_γ.

(viii) Local turbulent fluctuations: Both ET and SC models depend on the local turbulent dissipation/cascade rate (as well as e.g. magnetic field strengths). But, even on spatial scales resolved in our simulations, which are coherent on scales comparable to CR mean-free paths and scattering times, that rate has large (order-of-magnitude) local fluctuations on $\sim$0.1–100 pc scales. For example, if $\kappa \propto u^2$, where u is some local ISM property (like $|\delta v_{\rm turb}|$) that is lognormally distributed with factor $\sim$3 scatter, then the residence-time or scattering-weighted mean κ will be a factor $\sim$10 lower than the volume-weighted κ. This means that L_γ will generally be larger than assumed using just the 'median' properties of the ISM to estimate κ.

Clearly, one cannot fully capture these effects by post-processing CR transport in simple analytic or empirical galaxy models. The effects above produce the large systematic internal variations of κ and $v_{\rm st}$ in Figs 8 and 9. Moreover, almost all these effects go in the direction of increasing L_γ and CR confinement. They also explain why the required κ or $v_{\rm st}$ in our simulations are significantly larger than those obtained in 'leaky box' or flat halo diffusion models which assume free escape of $\sim$GeV protons outside of the thin or thick disc. They demonstrate why the connection between κ, L_γ, and $e_{\rm cr}$ in Figs 2–6 is not trivially linear as predicted by the toy model in Section 5.1.1.

5.1.4 Fast transport in neutral gas is insufficient

In some of our models $\tilde{\kappa}_{\rm eff}$ can be 'large' ($\tilde{\kappa}_{29} \gg 1$) in *neutral* gas, but relatively small in the ambient warm ionized gas (WIM and inner CGM). This is true by construction in our 'two-κ' model in Section 3.1.1, or due to ion-neutral damping in SC models. We saw in Section 4 that this reduces the predicted L_γ and collisional losses (and therefore the CR 'residence time' in the disc) by a surprisingly small amount (factor $<$2). There are two reasons for this. First, per Section 5.1.3 above, a neutral cloud or 'slab' of gas with local $\kappa_{\rm neutral} \rightarrow \infty$ will just converge to constant $e_{\rm cr}$ set by the 'boundary' condition in the ambient WIM, so if the WIM has low $\kappa_{\rm ion}$ and traps CRs, they will still spend time in the cold clouds inside that WIM. Second, even if we ignore the effect above and assume that the CR residence time in a local 'patch' simply scales with the local $\sim 1/\tilde{\kappa}_{\rm eff}$ (the 'free escape' limit), we note that L_γ and grammage scale with the hadronic losses as $L_\gamma \propto \int e_{\rm cr}\, \rho\, {\rm d}^3\boldsymbol{x} \propto \int (1/\kappa)\, {\rm d}M_{\rm gas} \propto M_{\rm ion}/\langle\tilde{\kappa}_{\rm ion}\rangle + M_{\rm neutral}/\langle\tilde{\kappa}_{\rm neutral}\rangle$ (where $M_{\rm ion}$ and $M_{\rm neutral}$ are the total mass of ionized gas and neutrals in the galaxy+CGM). So even if $\kappa_{\rm neutral} \rightarrow \infty$, this can only reduce L_γ by at most a factor $\sim 1 + M_{\rm neutral}/M_{\rm gas,\,total}$ relative to a model with $\tilde{\kappa} = \tilde{\kappa}_{\rm ion}$ everywhere. In dwarf galaxies, in particular the SMC, LMC, and M33, *most* of the gas is ionized, so this is a small correction, and even in the MW or M31, this is a factor only $\approx$1.5–2.

5.1.5 Can faster outflows or Alfvén speeds reduce the required transport coefficients?

It is clear that our 'advection+Alfvénic streaming' ($\kappa_{29} = 0$, $v_{\rm st} = v_A$) simulations severely overpredict the observed CR grammage, energy density, γ-ray luminosity, etc. However, large theoretical uncertainties remain in predicted galactic outflow and magnetic field properties (see e.g. Naab & Ostriker 2017, for a review). So although our comparisons between FIRE simulations and observations in previous work (see references in Section 1) suggests plausible agreement, it is possible that real galaxies feature significantly stronger outflows or magnetic fields, reducing the residence time even with $\kappa \rightarrow 0$. But even if we ignore all the complications described above, it seems implausible that this could significantly reduce the values of $\tilde{\kappa}_{\rm eff}$ or $\bar{v}_{\rm st,\,eff}$ required by the observations.

First consider outflows/pure advection: there are at least three major issues invoking outflows to provide 'most' of the CR transport. (1) The required outflow speeds must reach at least $\bar{v}_{\rm st,\,eff} \sim$ 300–3000 km s^{-1} (Fig. 10) in most galaxies at $z \sim 0$, including dwarfs – but these are much larger than outflow speeds observed in all but the most extreme starburst/AGN systems Martin (1999), Martin & Bouché (2009), Rupke (2018). (2) In the pure-advection limit, the 'residence time' of CRs is equivalent to the 'residence time' of gas in the ISM; but the observationally favoured CR residence times $\sim 10^7$ yr are much shorter than even a single Galactic dynamical time $\sim 10^8$ yr. So even gas which accretes falling in the free-fall velocity, mixes in a single dynamical time, and then accelerates outwards to the escape velocity will exceed observed CR residence times. (3) Most of the ISM observed (and simulated), even in dense star-forming regions, is *not in outflow* (Evans et al. 2009). Equivalently, most of the CR γ-ray emission, residence time, and grammage comes not from outflows but from the diffuse ISM; and of course the Galactic constraints on CRs (measured at Earth) come specifically from gas not in outflow. So even infinitely fast outflows will only reduce the required transport speeds in the non-outflowing ISM by a factor of order the ISM mass fraction in outflows (similar to our argument above regarding cold clouds), which is never larger than tens of percents.

Next, consider Alfvénic streaming. Here, the problem is obvious: to approach the required transport speeds and therefore observed grammage/residence times/γ-ray luminosities without invoking super-Alfvénic streaming or diffusion would require $v_A \sim v_{\rm st,\,eff} \sim 1000$ km s^{-1} in dwarf and MW-like galaxies, i.e. for typical ISM gas densities $n \sim 1$ cm^{-3} we would require $|\boldsymbol{B}| \sim 500$ μG, order of magnitude larger than observed. Even if we arbitrarily multiply the magnetic field strengths in our simulations by a factor $\sim$10 (larger than what is allowed by observations; see Guszejnov et al. 2020), we would still require a volume-weighted $v_{\rm st,\,eff} \gg v_A$ in Fig. 10.

It is therefore difficult if not impossible for these effects to alter our implied constraints on CR transport speeds by more than an order-unity factor.

5.2 Extrinsic turbulence

5.2.1 Alfvén modes

Consistent with conventional wisdom, we find that most standard ET models which assume scattering is dominated by resonant Alfvén waves modes (e.g. our 'Alfvén-C00' models and their variants, 'Alfvén-YL02', 'Alfvén-Hi', and related models) produce negligibly small CR scattering (i.e. higher κ) compared to the observationally inferred levels at $\sim$GeV energies (see Table 1). Correspondingly, these models alone (i.e. including no other scattering sources) underpredict the observed L_γ and MW grammage, as well as the CR energy density at the solar circle. Even if we neglect anisotropy and its effects on the scattering rate completely, giving $f_{\rm turb} = 1$ (our 'Alfvén-Max' model), this is only just barely able to reach the scattering levels observed.

5.2.2 Magnetosonic modes

If we assume a cascade of fast modes down to resonant scales $\sim r_{\rm L}$, assuming such modes are fully isotropic and ignoring any mode-damping (e.g. our 'Iso-K41' and 'Fast-NoDamp' models) then we would obtain excessively high scattering rates (low κ), clearly violating the observational constraints by factors of $\sim$10–100 (regardless of details of the power spectrum or whether we assume additional streaming at $\sim v_A$). But such models are clearly unphysical: in the warm WIM/CGM discussed above, accounting for just Braginskii viscosity as a damping mechanism and assuming trans-sonic turbulence, the equivalent Kolmogorov scale for fast (or perpendicular slow) modes is a factor $\ell_{\rm Kolm}/r_{\rm L} \sim 10^5\,(T/10^5\,K)^2$ larger than the gyro-resonant scales (in colder gas, ion-neutral damping and atomic/molecular collisional viscosity similarly gives $\ell_{\rm Kolm} \gtrsim 10^4\,r_{\rm L}$). Accounting for damping, the power in isotropic magnetosonic modes with wavelengths $\lambda \sim r_{\rm L}$ (hence their contribution to resonant scattering) should be vastly smaller than that in (undamped) Alfvén waves at similar wavelengths.

However, Yan & Lazarian (2004, 2008) argued that non-resonant fast modes with $\lambda \gg r_{\rm L}$ (plus undamped parallel gyro-resonant fast modes) can produce efficient CR scattering: we adopt their proposed scalings in our 'Fast-YL04' model and show that this could be allowed, and in fact could produce an order-unity fraction of the observed scattering in gas that is both fully ionized ($f_{\rm neutral} \lesssim 0.001$) and has $\beta \ll 1$. But this represents a small fraction of the ISM and almost none of the CGM, so likely contributes only modestly to observed scattering in total. Only by removing these restrictions ('Fast-Max') can this model approach the full observed scattering. We also caution that several assumptions in YL04 remain controversial including the degree of resonance-broadening, whether long-wavelength fast modes can efficiently scatter low-energy CRs via TTD, the $k^{-3/2}$ spectrum of the fast-mode power spectrum, and whether parallel fast modes follow the same spectrum below the scales where non-parallel modes are damped. Changing any of these decreases the implied scattering rate from fast modes by a large factor (e.g. our 'Fast-Mod' model).

5.3 Self-confinement

Again consistent with conventional wisdom, we find that 'standard' SC models predict much higher scattering rates and more efficient confinement of low-energy CRs compared to standard ET models (even the YL04 models). So it is reasonable to expect SC dominates over ET-induced scattering at $\sim$GeV. However, we actually find that 'default' or standard SC models predict *excessive* confinement – higher ν and lower κ, resulting in excessively high γ-ray luminosities, grammage, residence times, and CR energy densities – compared to observations. For reference, the predicted effective 'residence times' of CRs in 'SC:Default' model in MW-like haloes are $\gg 10^8$ yr, with CR energy densities $\gtrsim 10\,{\rm eV\,cm^{-3}}$, γ-ray production near the calorimetric limit, and grammage $X_s \gg 100\,{\rm g\,cm^{-2}}$. These characteristics are all in conflict with observations at the factor $\sim$10–1000 level.

As we discuss below, many of the model variations considered (see Table 1) do not resolve this issue: changing the CR energy by a factor $\sim$10, modest changes to the assumed turbulent structure, using equilibrium versus non-equilibrium treatments of CR transport, or adopting $v_A^{\rm ideal}$ or $v_A^{\rm ion}$ as the relevant Alfvén speed, all produce order-unity changes that are insufficient to explain these discrepancies. More fundamental changes, either invoking slower gyro-resonant growth rates (or lower scattering rates), or larger resonant-wave damping rates (or new damping mechanisms) by a factor $\sim$100, are required to reproduce the observations.

It is worth noting that in Table 1 and Figs 3 and 6, many of the observable predictions of the SC models appear to be almost 'bimodal'. Either the models predict excessive confinement near the calorimetric limit (with quite similar observables like those described above; e.g. our 'Default', '$\kappa \times 6$', '$v_A^{\rm ideal}$', '10 GeV', '$f_{\rm turb}-5$/DA', 'Non-Eqm', models), or they 'jump' to a new solution with much higher diffusivity, lower $L_\gamma/L_{\rm sf}$ and grammage, and lower $e_{\rm cr}$ at the MW solar circle, all in quite good agreement with the observations (e.g. our '$f_{\rm turb}-50/500/$K41', 'NE-$f_{\rm turb}-100$', '$f_{\rm QLT}-100$' models). This owes to the 'SC runaway' or 'bottleneck' effect described in Section 5.1.3: because SC models limit the absolute CR flux, the transport 'speed' (κ or $v_{\rm st}$) scales inversely with the CR energy density $e_{\rm cr}$ (equation 7). Thus, if there is a rapid injection of CRs (say from clustered SNe), $e_{\rm cr}$ rises rapidly, lowering κ, which slows CR escape, increasing $e_{\rm cr}$ and further lowering κ, in a runaway, until the CRs in that region lose their energy to collisions (hitting the calorimetric limit). To avoid this, the 'pre-factor' in the diffusive transport speeds, i.e. the damping rates $\Gamma_{\rm damp}$ or growth factor $f_{\rm QLT}$ must be large enough that CRs can efficiently escape these 'worst-case' (most efficiently trapped) environments. Once they do so, $e_{\rm cr}$ is made smooth by diffusion, and a 'smooth' or 'average' diffusivity becomes more reasonable.

5.3.1 Fast transport in neutral gas and choice of Alfvén speed

In the *neutral* ISM all the SC models here *do* predict large $\tilde{\kappa}_{\rm eff} \gg 10^{29}\,{\rm cm^2\,s^{-1}}$, regardless of how we treat the Alfvén speed when $f_{\rm ion} \ll 1$ (Section 2.4). If we take $v_A = v_A^{\rm ion} = f_{\rm ion}^{-1/2}\,v_A^{\rm ideal}$ in equation (7), then this becomes large for $f_{\rm ion} \ll 10^{-6}$ in GMCs, suppressing the '$\kappa_\parallel$' term in equation (7), but giving large $v_{\rm st} = v_A$ so $\kappa_{\rm eff} \sim \gamma_{\rm cr}\,v_{\rm st}\,\ell_{\rm cr} \sim 10^{31}\,{\rm cm^2\,s^{-1}}\,\ell_{\rm cr,\,kpc}\,B_{5\,\mu{\rm G}}\,n_{10}^{-1/2}\,(f_{\rm ion}/10^{-8})^{-1/2}$. If, instead, we take $v_A = v_A^{\rm ideal}$, then (taking $\Gamma \rightarrow \Gamma_{\rm in}$) we have $\kappa_{\rm eff} \sim \kappa_\parallel \sim 0.3 \times 10^{31}\,{\rm cm^2\,s^{-1}}\,\ell_{\rm cr,\,kpc}\,e_{\rm cr,\,eV}^{-1}\,n_{10}^{3/2}\,T_{1000}^{1/2}\,\gamma_{\rm L}$. But for the reasons discussed in Section 5.1.4 this alone does little to alter L_γ or the other observational constraints in Table 1 and Fig. 3: the overconfinement from SC models occurs in ionized, not neutral gas. And in the volume-filling WIM/CGM phases $f_{\rm ion} \sim 1$ and $v_A^{\rm ideal} \approx v_A^{\rm ion}$, so the choice of Alfvén speed does not produce any difference.

5.3.2 Equilibrium versus non-equilibrium models

We find that adopting the more detailed non-equilibrium evolution of the coefficients $\kappa_\parallel$, $v_{\rm st}$ as proposed in Thomas & Pfrommer (2019) (Section 3.3.2) makes little difference to our results, compared to adopting the 'local equilibrium' description in equation (7) (using the same damping coefficients). This is not surprising, as the time-scale for κ to reach the local equilibrium value is short $\sim \Gamma^{-1} \sim 3000\,{\rm yr}\,\Gamma_{-11}^{-1}$. In the non-equilibrium case, CRs do escape the galaxy *slightly* more easily, as they can 'free stream' a bit longer before e_A and the scattering rate 'build up'. However, this is likely at least somewhat artificially enhanced in our simulations here, because we adopt a 'reduced speed of light' $\tilde{c} < c$ (which increases the CR 'mean free path' $\sim \kappa/\tilde{c}$), so we caution against overinterpreting the result.

5.3.3 Overconfinement in the WIM and inner CGM

Consider our 'default' SC models (with $f_{\rm QLT} = f_{\rm cas} = 1$), in ionized gas representative of the warm and hot phases of the ISM and

CGM. Ion-neutral damping is negligible under these conditions.[17] Non-linear Landau (NLL) damping is also sub-dominant, and in fact *cannot* dominate $\Gamma_{\rm eff}$ in the WIM/inner CGM, without violating both the observational constraints on $e_{\rm cr}$ and $\tilde{\kappa}_{\rm eff}$: comparing $\Gamma_{\rm turb} + \Gamma_{\rm LL}$ (equations A2 and A3) and $\Gamma_{\rm NLL}$ (equation A4) in Appendix A, we see that $\Gamma_{\rm NLL} \gg (\Gamma_{\rm turb} + \Gamma_{\rm LL})$ requires $e_{\rm cr,\,eV} \gg 40\,(1+2.5/\beta^{1/2})^2\,\delta v_{10}^3\,n_1^2\,f_{\rm cas}^2\,T_4^{1/2}\,B_{\mu\rm G}^{-2}$. But if this condition were met, inserting these values of $e_{\rm cr}$ and $\Gamma_{\rm eff} \approx \Gamma_{\rm NLL}$ in equation (7) means the diffusivity would have to be *less than* $\kappa_\parallel \ll 5\times10^{25}\,{\rm cm^2\,s^{-1}}\,\ell_{\rm cr,\,kpc}^{1/2}\,\delta v_{10}^{-3/2}\,n_1^{-1/4}\,T_4^{1/2}$ (for any β), because $\kappa_\parallel$ for SC scales inversely with $e_{\rm cr}$. So in these environments $\Gamma_{\rm eff}$ is dominated by turbulent+linear Landau damping, which scale similarly as $\Gamma_{\rm LL} \approx 0.4\,\beta^{1/2}\,\Gamma_{\rm turb}$ and give $\kappa_\parallel \sim 10^{27}\,{\rm cm^2\,s^{-1}}\,(1+0.4\beta^{1/2})\,\delta v_{10}^{3/2}\,\ell_{\rm cr,\,kpc}\,\ell_{\rm turb,\,kpc}^{-1/2}\,n_1^{3/4}\,\gamma_{\rm L}^{1/2}\,e_{\rm cr,\,eV}^{-1}\,f_{\rm QLT}\,f_{\rm cas}$.

Although these values of κ and the $v_{\rm st} \approx v_A$ term[18] can become large in the *outer* CGM ($\gtrsim 30$ kpc, where $e_{\rm cr}$ is small, see Fig. 6), for $f_{\rm QLT}\,f_{\rm cas} \sim 1$ these are a factor of ∼30–300 smaller in the WIM/inner CGM than the values needed to explain observations (Table 1). As discussed above, it is also necessary in these models to overcome the SC runaway or bottleneck effect: this is particularly onerous in regions like superbubbles, which fill much of the volume around even new SNe (i.e. the CR sources, if SNe are clustered). With $n \sim 0.01$ and $e_{\rm cr,\,eV} \sim 10$ in these regions, the local $\tilde{\kappa}_{\rm eff}$ can be as low as $\sim 10^{24}\,{\rm cm^2\,s^{-1}}$ – equivalently the residence/escape time from an ~ 100 pc-size superbubble could reach ∼Gyr!

It is difficult to escape these conclusions: direct observational constraints on e.g. the turbulent velocity dispersions, scale-lengths, densities, and CR energy densities in the MW simply do not allow for large enough changes to those parameters to produce the required diffusivity without modifying $f_{\rm QLT}\,f_{\rm cas}$ above. The ISM parameters (e.g. n, T) are uncertain at the order-unity, not factor ∼100 level. The variations across different times in the galaxy history, and different galaxies like **m11f** and **m12i** (as well as other galaxies we have simulated described in Appendix D), fully span the 'allowed' observational range in these properties, and do not produce anywhere near the required values of L_γ or grammage with $f_{\rm QLT}\,f_{\rm cas} \sim 1$. And, even if the 'median' values of the scalings above for a given phase were promising, it is almost impossible to escape the conclusion that there will be substantial regions or local environments in the MW where the particular $\kappa_{\rm eff}$ predicted above would be very low, producing a severe 'bottleneck' unless, again, $f_{\rm QLT}\,f_{\rm cas}$ or some related factor can be made factor ∼100 larger.

5.3.4 *Possible resolutions*

Reconciling SC models with observations fundamentally requires factor ∼100 lower scattering rates ν (and correspondingly larger $\tilde{\kappa}_{\rm eff}$) in the WIM/inner CGM, compared to the predictions obtained with the most commonly assumed scalings (our 'default' model). Qualitatively, there could be several explanations for the discrepancy:

(i) **Inefficient scattering:** If CR scattering by gyro-resonant waves is much weaker than usually assumed[19] (for the same $\delta\boldsymbol{B}[r_{\rm L}]$ or e_A), this would directly lower ν. Gyro-resonant waves have a reasonably well-understood structure (see e.g. Zirakashvili, Ptuskin & Völk 2008; Ohira et al. 2009; Riquelme & Spitkovsky 2009) and the amplitudes predicted here are generally modest (for diffusivity $\tilde{\kappa}_{29}$, the gyro-resonant $|\delta\boldsymbol{B}[r_{\rm L}]|/|\boldsymbol{B}| \sim 3\times10^{-4}\,(\gamma_{\rm L}/B_{\mu\rm G}\,\tilde{\kappa}_{29})^{1/2}$); however, two recent works studying the saturation of the gyro-resonant instability using the PIC method suggest possible ways that the effective ν might be lower than the QLT prediction. First, Bai et al. (2019) find that the time required for the CR distribution to become fully isotropic in the Alfvén-wave frame is much longer than predicted by the QLT estimate. This behaviour arises because of particularly inefficient scattering across the zero pitch angle ($\mu = 0$) barrier, which is both slow and requires scatterers of very short wavelength compared to $r_{\rm L}$ (Völk 1973). Secondly, in the highly anisotropic regime most relevant to regions close to sources, Holcomb & Spitkovsky (2019) find very inefficient saturation of the gyro-resonant instability even when the self-excited Alfvén waves reach very large amplitudes, because only a single helicity (handedness) of Alfvén wave is produced by the CRs. Such an effect may help to limit the SC 'runaway' (see Sections 5.1.3 and 5.3.3) in regions with high $e_{\rm cr}$.

(ii) **Lower gyro-resonant growth rates:** If the growth rate of the gyro-resonant instability is a factor $f_{\rm QLT}^{-1}$ smaller compared to the usual linear-theory expression $\Gamma_{\rm grow}^{\rm linear} \sim \Omega\,(n_{\rm cr}/n_i)\,(\bar{v}_{\rm st}/v_A - 1)$, then the quasi-linear saturation amplitude of $\nu \rightarrow \nu/f_{\rm QLT}$ (and $\kappa \rightarrow f_{\rm QLT}\,\kappa$). In the WIM/CGM, we have $\beta \gg 1$, $e_{\rm cr}/e_{\rm B} \gg 1$, $\bar{v}_{\rm st}/v_A \sim 300$–$1000 \gg 1$, regimes where the instability is not well studied and could potentially be strongly modified.[20] The results of Bai et al. (2019) may again be of interest, if smaller scale modes excited by low-μ and lower energy particles are required to fully saturate the gyro-resonant instability. Since such particles are much less numerous, implying the growth rate of the resonant modes is lower, the damping-growth balance that is usually assumed to saturate the instability and determine κ (see Section 3.3) might occur at significantly lower Alfvén-wave amplitudes than usually assumed. It seems plausible that such an effect could lead to significant enhancements in the SC diffusion rates, although clearly more work is needed.

(iii) **Larger damping rates or alternative mechanisms:** Since the saturation amplitude of $|\delta\boldsymbol{B}[r_{\rm L}]|^2/|\boldsymbol{B}|^2$, hence scattering rates, are inversely proportional to the damping rate $\Gamma_{\rm eff}$ in the quasi-linear theory models considered here (giving $\kappa \propto \Gamma_{\rm eff}$), it may instead be that damping rates are underestimated. We stress that the required damping rates are still very small in absolute terms: $\Gamma_{\rm damp} \gtrsim 10^{-7}\,\Omega$ gives the required $\tilde{\kappa}_{29} \gtrsim 1$. Also, as discussed above, any such damping must operate efficiently in the ionized ISM and inner CGM:

[17] While ion-neutral damping is efficient in dense gas ($n_1 \gg 1$) as $f_{\rm neutral} \rightarrow 1$ (with $f_{\rm ion} \lesssim 10^{-6}$ very small), if $f_{\rm neutral} \lesssim 1$ (so $f_{\rm ion}$ is not $\ll$1), then achieving an effective isotropic diffusivity $\tilde{\kappa}_{29} \gtrsim 1$ requires $f_{\rm neutral} \gtrsim e_{\rm cr,\,eV}/(\ell_{\rm cr,\,kpc}\,n_1^{3/2})$. So at densities $n \lesssim 1\,{\rm cm^{-3}}$, or temperatures $T \gtrsim 2\times10^4$ K (where $f_{\rm neutral} \ll 0.01$ drops exponentially), $\Gamma_{\rm IN}$ is small both compared to other damping mechanisms ($\Gamma_{\rm IN} \ll \Gamma_{\rm turb} + \Gamma_{\rm LL} + \Gamma_{\rm NLL}$) and compared to the observationally required damping rates.

[18] For $v_{\rm st} = v_A$, the corresponding $\kappa_{\rm eff} \sim \gamma_{\rm cr}\,v_{\rm st}\,\ell_{\rm cr} \sim 10^{27}\,{\rm cm^2\,s^{-1}}\,B_{\mu G}\,\ell_{\rm cr,\,kpc}\,n_1^{-1}$.

[19] Uniformly decreasing the predicted scattering rate ν by a factor $f_{\rm scatter}$, all else equal, in our models, is equivalent to multiplying $\kappa_\pm$ given by the closure-relation in equation (B3) by $f_{\rm scatter}$, which in turn multiplies the 'local equilibrium' $\kappa_\parallel$ in equation (7) by $f_{\rm scatter}$ as well, exactly identical to our '$f_{\rm QLT}$' parameter.

[20] For the conditions of interest in the WIM/CGM and $\tilde{\kappa}_{29} \sim 1$, we expect large $\beta \sim 35\,n_1\,T_4\,B_{\mu G}^{-2} \gg 1$ (using our standard notation to scale T relative to 10^4 K, etc.), large ratio of CR to magnetic energy $e_{\rm cr}/e_{\rm B} \sim 40\,e_{\rm cr,\,eV}\,B_{\mu G}^{-2} \gg 1$, small fractional magnetic fluctuations at the gyro scale $|\delta\boldsymbol{B}[r_{\rm L}]|/|\boldsymbol{B}| \sim 3\times10^{-4}\,(\gamma_{\rm L}/B_{\mu\rm G}\,\tilde{\kappa}_{29})^{1/2} \ll 1$, small CR number density relative to ions $n_{\rm cr}/n_i \sim 10^{-9}\,e_{\rm cr,\,eV}\,n_1^{-1}\,\gamma_{\rm L}^{-1} \ll 1$, and large ratio of 'effective' streaming speed to Alfvén speed (corresponding to this diffusivity) $\bar{v}_{\rm st}/v_A \sim 300\,\tilde{\kappa}_{29}\,n_1^{1/2}\,B_{\mu\rm G}^{-1}\,\ell_{\rm cr,\,kpc}^{-1} \gg 1$.

ion-neutral damping is efficient where neutral fractions are large but does not resolve the transport bottlenecks that appear in the fully ionized WIM/HIM and inner CGM.

One possibility is that the turbulent (or linear Landau) damping rates are larger by a factor ∼100; i.e. the turbulent dissipation or cascade time $t_{\rm cascade}$ is shorter by a factor $f_{\rm cas}\sim 100$ at resonant scales. This may appear to be a large factor, but recall that the cascade models used to infer $t_{\rm cascade}$ and $\Gamma_{\rm turb}$ are extrapolated by factors reaching $\sim 10^{8}-10^{10}$ in scale from the ISM/CGM driving scales to $\sim r_{\rm L}$, so even quite small changes to the structure of the cascade could produce such a factor (although at least some of the variations we consider actually change this with the wrong sign, giving lower $\Gamma_{\rm turb}$). If other mechanisms (unresolved here), could directly drive turbulence on small scales (with e.g. an isotropic dispersion of $\sim 0.1\,{\rm km\,s^{-1}}$ on scales $\sim r_{\rm L}$) this would also resolve the discrepancy. And even given a particular cascade, we caution that the standard Farmer & Goldreich (2004) model for how such a cascade damps resonant Alfvén waves has a number of uncertainties. Further, it remains untested in non-linear simulations.

There could also be additional damping/saturation mechanisms for gyro-resonant instabilities, not considered in our default models: e.g. non-linear effects, or self-interactions, or parasitic modes involving other (non-resonant) instabilities. There are many linear instabilities that couple magnetic fields, acoustic modes, gas, and other plasma components on scales $\sim r_{\rm L}$. For example, the acoustic instabilities studied in Drury & Falle (1986), Begelman & Zweibel (1994), Kempski et al. (2019) could be significant precisely in the warm/hot ionized medium when CR pressure gradients are weak. Recently Squire & Hopkins (2018a, 2018b), Hopkins, Squire & Seligman (2020c) discovered a class of 'resonant drag instabilities' (RDIs) between dust and gas or magnetic fields that includes a sub-family of 'Alfvén RDIs' and 'CR-like' RDIs which directly interact with Alfvén waves and are unstable at wavelengths $\sim r_{\rm L}$ in the WIM with growth rates (for $\sim 0.1\,\mu$m grains) $\Gamma_{\rm RDI}\gg 10^{-11}\,{\rm s^{-1}}$, making them also potentially interesting here.

6 COMPARISON TO OTHER COSMOLOGICAL SIMULATIONS AND PREVIOUS WORK

To our knowledge, there has been no previous work comparing the various ET or SC-motivated CR transport models above in galaxy formation simulations. Considering 'constant-diffusivity' models, outside of Papers I and II, only a few other studies have compared galaxy simulations with CR transport to the any of the observables discussed here. Salem et al. (2016) considered 'constant-diffusivity' models without MHD or hadronic losses, with isotropic $\tilde{\kappa}_{\rm eff,\,29}\sim 0.03$–$0.3$ ($v_{\rm st}=0$), arguing that higher diffusivities are needed to match diffuse γ-ray emission constraints. Pfrommer et al. (2017) and Buck et al. (2020) considered anisotropic MHD simulations with $v_{\rm st}=0$, and $\kappa_{\parallel}=0$ or $\kappa_{\parallel}=10^{28}\,{\rm cm^2\,s^{-1}}$ (i.e. $\kappa_{29}=0.1$). They concluded that with these low-$\kappa_{\parallel}$ values, almost all galaxies produce L_{γ} within a factor ∼1−3 of the calorimetric limit, with grammage $X_s\gtrsim 100\,{\rm g\,cm^{-2}}$ in MW-like galaxies (see Appendix E), and $e_{\rm cr}\sim 20\,{\rm eV\,cm^{-3}}$ at the 'solar circle'. All of these results are similar to our constant-diffusivity models with similar $\kappa_{\parallel}$, supporting our conclusions regarding both the transport speeds required and the relatively minor effect from dense gas. However, Buck et al. (2020) argue that their low-$\kappa_{\parallel}$ models, even their 'advection only' models ($v_{\rm st}=0,\kappa_{\parallel}=0$), can reproduce the γ-ray observations (and therefore disagreed with our Paper I conclusions). We discuss this in detail in Appendix E, arguing that the discrepancy stems not from a theoretical or simulation difference, but from how the γ-ray observations of the SMC/LMC/M33/MW/M31 are plotted, as well as their neglect of MW grammage and energy–density constraints.

Within the MW, there is a long history of modelling CR transport in simplified analytic, time-static, smooth 'disc+halo' models (generally neglecting phase structure or magnetic fields/anisotropy, but see e.g. Blasi & Amato 2012b), again almost exclusively with 'constant-diffusivity' models (although a few studies have considered models where κ varies with e.g. galactocentric radius in some idealized fashion; see Liu, Yao & Guo 2018). As we noted above and in Papers I and II, our favoured values of $\kappa_{\parallel}$ and the scalings in e.g. Section 5.1.1 for our constant-$\kappa_{\parallel}$ models are broadly consistent with these studies (compare Blasi & Amato 2012a; Vladimirov et al. 2012; Gaggero et al. 2015; Cummings et al. 2016; Guo et al. 2016; Jóhannesson et al. 2016; Korsmeier & Cuoco 2016; Evoli et al. 2017; Amato & Blasi 2018), *if* we compare to MW models that include an extended (∼ 10 kpc) gaseous halo, and account for the difference between the isotropically averaged diffusivity $\tilde{\kappa}_{\rm eff}$ usually measured in those models and the parallel $\kappa_{\rm eff}$ (a factor of ∼3 larger) defined here. These analytic constant-κ models generally find $\kappa_{29}\sim 1$ required to reproduce the observations: a factor ∼10–100 larger than the diffusivity implied by older models that ignored any halo and assumed CRs escape outside the thin-disc scale height (∼ 200 pc).

7 CONCLUSIONS

We have presented the first numerical simulations that simultaneously follow self-consistent cosmological galaxy formation with CGM and ISM phase structure coupled to explicit physically motivated dynamical models of low-energy (∼ GeV) CR transport, where the relevant transport parameters (effective diffusivity κ and/or streaming speed $v_{\rm st}$) are functions of the local plasma properties. We consider a wide range of micro-physical CR transport models, motivated by ET and SC scenarios, and compare the results of these directly to observational constraints in the MW and from nearby galaxies including γ-ray emission, CR energy densities, grammage, and residence times. We show that this is able to strongly constrain or rule out a large variety of proposed models and scalings for κ and $v_{\rm st}$. Our major conclusions include:

(i) **The 'leaky box' (or 'flat halo diffusion') is a bad approximation, and the CGM gas is critical**: In *all* physically motivated models we consider, CRs below $\lesssim 10$ GeV remain confined (mean-free-paths $\lambda_{\rm mfp}\ll r$) at all galactocentric radii out to well past the virial radius (scales $\lesssim$Mpc), even though κ tends to increase slowly with radius. This implies that the CR scattering and confinement is strongly influenced by the presence of extended gaseous haloes in the CGM (which are ubiquitous and contain *most* of the baryons) with scale-lengths ∼ 10–50 kpc. 'Toy' or analytic CR transport models must include such large, *continuous* haloes when considering ∼ GeV CRs. This, in turn, necessarily implies larger transport speeds, compared to simpler leaky-box or flat-halo diffusion models.

(ii) **There is no 'single' diffusivity, and ISM/CGM phase structure is important:** Also in all the physically motivated models here, CR transport parameters (κ, $v_{\rm st}$) depend strongly on properties like the local turbulent dissipation rate, magnetic field strength, ionization fraction, and gas density, which vary by *orders of magnitude* locally in both time and space along the trajectories of individual CRs owing to, e.g. rapidly time-varying ISM phase structure. Because of these variations, even taking spatial-and-time averages within a specific galactocentric annulus, there is no 'single' mean κ (or $v_{\rm st}$). The volume-weighted and 'residence time' or 'scattering rate'-weighted κ (or $v_{\rm st}/v_A$) can differ by factors ∼10–100.

(iii) **Relatively 'large' transport speeds are required:** In any models considered which reproduce the observational constraints, the effective *scattering-rate-weighted* mean parallel diffusivity $\kappa_{\rm eff,\,\parallel} \sim 10^{29}$–$10^{31}\,{\rm cm^2\,s^{-1}}$ in the ISM of dwarf and $\sim L_*$ galaxies within $\lesssim 10$ kpc. This $\kappa_{\rm eff,\,\parallel}$ typically rises by factors $\sim$10–100 in the CGM from $\sim$30 to 300 kpc. It also varies systematically between galaxies (becoming somewhat larger in smaller dwarfs) and with redshift (decreasing, on average, at high-z). This corresponds to highly super-Alfvénic streaming, with bulk transport speed $\bar{v}_{\rm st} \sim 10$–$1000\,v_A$. If one accounts for large CGM haloes, fluctuations in local ISM properties, and isotropic versus anisotropic diffusion, this required diffusivity is consistent with simple analytic and idealized models, but we emphasize that almost all non-linear effects in our simulations tend to enhance CR confinement (increasing the required $\kappa_{\rm eff,\,\parallel}$).

(iv) **Fast CR transport in neutral gas *alone* is not enough:** Neutral (molecular or H I) gas clouds in the ISM are embedded in volume-filling WIM and hotter ionized gas, most of which has local neutral fractions $1 - f_{\rm ion} \lesssim 0.01$. The entire galaxy is itself embedded in 'inner CGM' (scales $\lesssim$ 10–30 kpc) gas with densities $n \sim 10^{-3}$–$0.1\,{\rm cm^{-3}}$, temperatures $\sim 3\times10^4$–10^6 K, and $1 - f_{\rm ion} \ll 0.01$. So even if $\kappa \rightarrow \infty$ in neutral gas, CRs simply reach a constant energy density inside cold/neutral clouds, with their energy density and transport speed rate-limited by the boundary condition of this ionized 'cocoon'.

(v) **Extrinsic turbulence (probably) does not dominate:** As widely assumed, most physically motivated ET models predict lower scattering rates for $\sim$GeV CRs, compared to what is observed (indicating that ET does not dominate $\sim$GeV CR scattering). However, if we ignore anisotropy and damping (e.g. assume an isotropic Kolmogorov turbulent spectrum from the driving scale $\ell_{\rm turb} \sim 0.1$ kpc down to the gyro scale $r_{\rm L} \sim 0.1$ au), the scattering rate from ET alone would severely exceed observational limits. Interestingly, one particular version of the proposed model from Yan & Lazarian (2004) for scattering by fast modes with wavelengths $\gg r_{\rm L}$ could produce scattering rates similar to SC in gas which is fully ionized and also has $\beta \ll 1$, but this represents a small fraction of the ISM/CGM and the assumptions made in that model remain highly uncertain. Moreover, once anisotropy and damping are accounted for, *all* ET models considered here predict the incorrect qualitative dependence of grammage/residence time on rigidity at energies $\sim$GeV-TeV (opposite the observed trend, regardless of the turbulent spectrum).

(vi) **Simple quasi-linear expectations for self-confinement produce excessive confinement:** Using the most common quasi-linear estimates for CR transport governed by SC – i.e. assuming scattering rates are set by resonant Alfvén-wave energy densities that are themselves set by the competition between gyro-resonant streaming instability growth and damping with standard literature estimates for turbulent, ion-neutral, and Landau damping rates – we predict galaxy-integrated scattering rates that are a factor $\sim$100 larger than observationally allowed. This primarily comes from the volume-filling WIM and 'inner CGM' discussed above, where ion-neutral damping is negligible (transport is fast, in these models, in neutral gas). We discuss possible resolutions in Section 5.3.4. It is plausible that scattering caused by the gyro-resonant instability could be less efficient than naive (quasi-)linear theory expectations by a factor $f_{\rm cas} \sim 100$; for example, due to inefficient isotropization of the CR distribution function across small pitch angles (Bai et al. 2019), or because near-source scattering is weaker than expected (Holcomb & Spitkovsky 2019). Alternatively, damping rates from turbulence or linear-Landau effects could be larger by a factor $f_{\rm cas} \sim 100$, if the turbulence is less-strongly anisotropic (as compared what is implied by usual critical-balance arguments), or if there are processes which can directly drive turbulence on scales closer to $r_{\rm L}$. It is also possible that different damping processes, not usually considered, could dominate in the fully ionized, warm, intermediate density environments that are particularly important for global CR transport.

(vii) **Models exist which can reproduce CR observations:** We emphasize that if we lower the 'default' SC scattering rate by a factor $f_{\rm QLT}$ or $f_{\rm cas} \sim 100$, then this model simultaneously reproduces (from fully cosmological simulations) *all* the observational constraints we consider, including γ-ray measurements from SMC/LMC/M33/MW/M31 through starburst galaxies, the observed CR energy density at the solar circle, MW grammage and residence times *and* their dependence on rigidity. That this is possible at all, with just one dimensionless normalization constant ($f_{\rm QLT}\,f_{\rm cas}$) set to a single universal value, is extremely encouraging. We can also reproduce these observations at $\sim$ 1 GeV with a constant-κ model if we set $\kappa_{29} \sim 3$–30, or with a scaling motivated by ET if we artificially increase the ET scattering rate with e.g. our 'Alfvén-Max' or 'Fast-Max' models, although neither the constant-κ model nor these variant ET models predict the observed dependence of grammage/residence time on rigidity (as the SC-motivated models do).

Our goal in this study is primarily to place first observational constraints on various 'a priori' models which have been proposed in the literature for how the effective CR transport parameters (parallel diffusivity and/or streaming/drift speeds) depend on local plasma properties. We emphasize that our resolution is nowhere near sufficient to *predict* these scalings: rather we implement fully dynamical CR evolution using different scalings derived from analytic models or PIC simulations. The qualitatively important resolution criteria are that we begin to resolve the multiphase structure within the ISM and CGM (which determines these scalings) and that we at least marginally resolve the deflection length of CRs (so their trajectories through that medium can be followed). Our hope is that the conclusions above motivate some general conclusions for galaxy-scale CR transport, and motivate additional theoretical work exploring CR transport in SC scenarios and/or fast-mode scattering. The simulations are of course an imperfect representation of reality: we discuss a wide range of additional caveats in Appendix D, including resolution, numerical implementation details, form of the CR flux equation, equilibrium versus non-equilibrium treatments, statistics (simulating additional galaxies), explicit inclusion of perpendicular diffusivities, and more. The uncertainties owing to some of these choices can be significant for some predictions (for extensive discussion of how resolution influences the ISM structure itself, see e.g. Hopkins et al. 2018a, b), but for our purposes here they generally produce factor $\lesssim$2 differences in the predicted γ-ray luminosity or grammage given a *fixed* physical model for CR transport (see Papers I and II). In contrast, different choices of CR transport models produce factor $\gg$1000 differences. Given that the most interesting conclusions discussed above are factor $\sim$100-level effects, it is likely that our conclusions are robust to these and other order-unity effects.

ACKNOWLEDGEMENTS

We thank the anonymous referee for helpful suggestions. Support for PFH was provided by NSF Collaborative Research Grants 1715847 and 1911233, NSF CAREER grant 1455342, NASA grants 80NSSC18K0562, and JPL 1589742. CAFG was supported by NSF 1517491, 1715216, and CAREER 1652522; NASA 17-ATP17-0067; and by a Cottrell Scholar Award. DK was supported by NSF grant AST-1715101 and the Cottrell Scholar Award. Numerical

calculations were run on the Caltech compute cluster 'Wheeler', allocations from XSEDE TG-AST130039 and PRAC NSF.1455342 supported by the NSF, and NASA HEC SMD-16-7592. Data used in this work were hosted on facilities supported by the Scientific Computing Core at the Flatiron Institute, a division of the Simons Foundation.

DATA AVAILABILITY STATEMENT

The data supporting the plots within this article are available on reasonable request to the corresponding author. A public version of the GIZMO code is available at http://www.tapir.caltech.edu/~phopkins/Site/GIZMO.html. Additional data including simulation snapshots, initial conditions, and derived data products are available at http://fire.northwestern.edu.

REFERENCES

Aguilar M. et al., 2016, Phys. Rev. Lett., 117, 091103
Aguilar M. et al., 2018, Phys. Rev. Lett., 120, 021101
Amato E., Blasi P., 2018, Adv. Space Res., 62, 2731
Bai X.-N., Caprioli D., Sironi L., Spitkovsky A., 2015, ApJ, 809, 55
Bai X.-N., Ostriker E. C., Plotnikov I., Stone J. M., 2019, ApJ, 876, 60
Begelman M. C., Zweibel E. G., 1994, ApJ, 431, 689
Bell A. R., 2004, MNRAS, 353, 550
Benincasa S. M. et al., 2020, MNRAS, 497, 3993
Blasi P., 2017, MNRAS, 471, 1662
Blasi P., Amato E., 2012a, J. Cosmol. Astropart. Phys., 2012, 010
Blasi P., Amato E., 2012b, J. Cosmol. Astropart. Phys., 2012, 011
Boldyrev S., 2006, Phys. Rev. Lett., 96, 115002
Boldyrev S., Nordlund Å., Padoan P., 2002, ApJ, 573, 678
Buck T., Pfrommer C., Pakmor R., Grand R. J. J., Springel V., 2020, MNRAS, 497, 1712
Burgers J., 1973, The Nonlinear Diffusion Equation: Asymptotic Solutions and Statistical Problems. D. Reidel Pub. Co., Springer, The Netherlands
Burkhart B., Falceta-Gonçalves D., Kowal G., Lazarian A., 2009, ApJ, 693, 250
Butsky I. S., Quinn T. R., 2018, ApJ, 868, 108
Chan T. K., Kereš D., Hopkins P. F., Quataert E., Su K. Y., Hayward C. C., Faucher-Giguère C. A., 2019, MNRAS, 488, 3716(Paper I)
Chandran B. D. G., 2000, Phys. Rev. Lett., 85, 4656
Chen C. H. K., 2016, J. Plasma Phys., 82, 535820602
Cho J., Lazarian A., 2003, MNRAS, 345, 325
Colbrook M. J., Ma X., Hopkins P. F., Squire J., 2017, MNRAS, 467, 2421
Cowsik R., Burch B., Madziwa-Nussinov T., 2014, ApJ, 786, 124
Cummings A. C. et al., 2016, ApJ, 831, 18
Dorfi E. A., Breitschwerdt D., 2012, A&A, 540, A77
Drury L. O., Falle S. A. E. G., 1986, MNRAS, 223, 353
El-Badry K. et al., 2018a, MNRAS, 473, 1930
El-Badry K. et al., 2018b, MNRAS, 477, 1536
Enßlin T. A., Pfrommer C., Springel V., Jubelgas M., 2007, A&A, 473, 41
Enßlin T. A., Pfrommer C., Miniati F., Subramanian K., 2011, A&A, 527, A99
Escala I. et al., 2018, MNRAS, 474, 2194
Evans N. J. et al., 2009, ApJS, 181, 321
Everett J. E., Zweibel E. G., Benjamin R. A., McCammon D., Rocks L., Gallagher III J. S., 2008, ApJ, 674, 258
Evoli C., Gaggero D., Vittino A., Di Bernardo G., Di Mauro M., Ligorini A., Ullio P., Grasso D., 2017, J. Cosmol. Astropart. Phys., 2017, 015
Farber R., Ruszkowski M., Yang H.-Y. K., Zweibel E. G., 2018, ApJ, 856, 112
Farmer A. J., Goldreich P., 2004, ApJ, 604, 671
Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, ApJ, 703, 1416
Fu L., Xia Z. Q., Shen Z. Q., 2017, MNRAS, 471, 1737
Gaggero D., Urbano A., Valli M., Ullio P., 2015, Phys. Rev. D, 91, 083012
Garrison-Kimmel S. et al., 2018, MNRAS, 481, 4133
Garrison-Kimmel S. et al., 2019, MNRAS, 487, 1380
Genel S., Vogelsberger M., Nelson D., Sijacki D., Springel V., Hernquist L., 2013, MNRAS, 435, 1426
Girichidis P., Naab T., Hanasz M., Walch S., 2018, MNRAS, 479, 3042
Goldreich P., Sridhar S., 1995, ApJ, 438, 763
Grand R. J. J. et al., 2017, MNRAS, 467, 179
Griffin R. D., Dai X., Thompson T. A., 2016, ApJ, 823, L17
Grudić M. Y., Hopkins P. F., Faucher-Giguère C.-A., Quataert E., Murray N., Kereš D., 2018, MNRAS, 475, 3511
Grudić M. Y., Hopkins P. F., Lee E. J., Murray N., Faucher-Giguère C.-A., Johnson L. C., 2019, MNRAS, 488, 1501
Guo F., Oh S. P., 2008, MNRAS, 384, 251
Guo Y.-Q., Tian Z., Jin C., 2016, ApJ, 819, 54
Guszejnov D., Hopkins P. F., Ma X., 2017, MNRAS, 472, 2107
Guszejnov D., Hopkins P. F., Graus A. S., 2019, MNRAS, 485, 4852
Guszejnov D., Grudić M. Y., Offner S. S. R., Boylan-Kolchin M., Faucher-Gigère C.-A., Wetzel A., Benincasa S. M., Loebman S., 2020, MNRAS, 492, 488
Haberl F., Sturm R., 2016, A&A, 586, A81
Hafen Z. et al., 2019, MNRAS, 488, 1248
Hagen L. M. Z., Siegel M. H., Hoversten E. A., Gronwall C., Immler S., Hagen A., 2017, MNRAS, 466, 4540
Hall M., Courteau S., Dutton A. A., McDonald M., Zhu Y., 2012, MNRAS, 425, 2741
Harris J., Zaritsky D., 2004, AJ, 127, 1531
Hayward C. C., Hopkins P. F., 2017, MNRAS, 465, 1682
Holcomb C., Spitkovsky A., 2019, ApJ, 882, 3
Holman G. D., Ionson J. A., Scott J. S., 1979, ApJ, 228, 576
Hony S. et al., 2015, MNRAS, 448, 1847
Hopkins P. F., 2013, MNRAS, 430, 1880
Hopkins P. F., 2015, MNRAS, 450, 53
Hopkins P. F., 2016, MNRAS, 462, 576
Hopkins P. F., 2017, MNRAS, 466, 3387
Hopkins P. F., Raives M. J., 2016, MNRAS, 455, 51
Hopkins P. F., Hernquist L., Martini P., Cox T. J., Robertson B., Di Matteo T., Springel V., 2005, ApJ, 625, L71
Hopkins P. F., Quataert E., Murray N., 2012a, MNRAS, 421, 3488
Hopkins P. F., Quataert E., Murray N., 2012b, MNRAS, 421, 3522
Hopkins P. F., Narayanan D., Murray N., 2013a, MNRAS, 432, 2647
Hopkins P. F., Kereš D., Murray N., Hernquist L., Narayanan D., Hayward C. C., 2013b, MNRAS, 433, 78
Hopkins P. F., Keres D., Onorbe J., Faucher-Giguere C.-A., Quataert E., Murray N., Bullock J. S., 2014, MNRAS, 445, 581
Hopkins P. F. et al., 2018a, MNRAS, 477, 1578
Hopkins P. F. et al., 2018b, MNRAS, 480, 800
Hopkins P. F., Grudić M. Y., Wetzel A., Kereš D., Faucher-Giguère C.-A., Ma X., Murray N., Butcher N., 2020a, MNRAS, 491, 3702
Hopkins P. F. et al., 2020b, MNRAS, 492, 2465(Paper II)
Hopkins P. F., Squire J., Seligman D., 2020c, MNRAS, 496, 2123
Indriolo N., McCall B. J., 2012, ApJ, 745, 91
Indu G., Subramaniam A., 2011, A&A, 535, A115
Iroshnikov P. S., 1963, Astron. Zh., 40, 742
Ji S. et al., 2020, MNRAS, 496, 4221
Jiang Y.-F., Oh S. P., 2018, ApJ, 854, 5
Jóhannesson G. et al., 2016, ApJ, 824, 16
Jokipii J. R., 1966, ApJ, 146, 480
Jubelgas M., Springel V., Enßlin T., Pfrommer C., 2008, A&A, 481, 33
Kachelrieß M., Semikoz D. V., 2019, Prog. Part. Nucl. Phys., 109, 103710
Keating L. C. et al., 2020, MNRAS, 499, 837
Kempski P., Quataert E., Squire J., Kunz M. W., 2019, MNRAS, 486, 4013
Kennicutt R. C., Jr, 1998, ApJ, 498, 541
Kewley L. J., Ellison S. L., 2008, ApJ, 681, 1183
Kolmogorov A., 1941, Akad. Nauk SSSR Dokl., 30, 301
Korsmeier M., Cuoco A., 2016, Phys. Rev. D, 94, 123019
Kraichnan R. H., 1965, Phys. Fluids, 8, 1385
Kritsuk A. G., Norman M. L., Padoan P., Wagner R., 2007, ApJ, 665, 416

Kulsrud R. M., 2005, Plasma Physics for Astrophysics. Princeton Univ. Press, Princeton, NJ
Kulsrud R. M., Pearce W. P., 1969, ApJ, 156, 445
Lacki B. C., Thompson T. A., Quataert E., Loeb A., Waxman E., 2011, ApJ, 734, 107
Lazarian A., 2016, ApJ, 833, 131
Leahy D. A., 2017, ApJ, 837, 36
Leitherer C. et al., 1999, ApJS, 123, 3
Liu W., Yao Y.-h., Guo Y.-Q., 2018, ApJ, 869, 176
Lopez L. A., Auchettl K., Linden T., Bolatto A. D., Thompson T. A., Ramirez-Ruiz E., 2018, ApJ, 867, 44
McKenzie J. F., Voelk H. J., 1982, A&A, 116, 191
Maggi P. et al., 2019, A&A, 631, A127
Mannheim K., Schlickeiser R., 1994, A&A, 286, 983
Mao S. A., Ostriker E. C., 2018, ApJ, 854, 89
Maoz D., Badenes C., 2010, MNRAS, 407, 1314
Martin C. L., 1999, ApJ, 513, 156
Martin C. L., Bouché N., 2009, ApJ, 703, 1394
Maurin D., Putze A., Derome L., 2010, A&A, 516, A67
Montgomery D., Turner L., 1981, Phys. Fluids, 24, 825
Muratov A. L., Kereš D., Faucher-Giguère C.-A., Hopkins P. F., Quataert E., Murray N., 2015, MNRAS, 454, 2691
Naab T., Ostriker J. P., 2017, ARA&A, 55, 59
Noël N. E. D., Aparicio A., Gallart C., Hidalgo S. L., Costa E., Méndez R. A., 2009, ApJ, 705, 1260
Ohira Y., Reville B., Kirk J. G., Takahara F., 2009, ApJ, 698, 445
Orr M. E. et al., 2018, MNRAS, 478, 3653
Orr M. E., Hayward C. C., Hopkins P. F., 2019, MNRAS, 486, 4724
Padovani M., Galli D., Glassgold A. E., 2009, A&A, 501, 619
Pakmor R., Pfrommer C., Simpson C. M., Springel V., 2016, ApJ, 824, L30
Pan L., Padoan P., Kritsuk A. G., 2009, Phys. Rev. Lett., 102, 034501
Pfrommer C., Pakmor R., Simpson C. M., Springel V., 2017, ApJ, 847, L13
Ptuskin V. S., Moskalenko I. V., Jones F. C., Strong A. W., Zirakashvili V. N., 2006, ApJ, 642, 902
Putze A., Derome L., Maurin D., 2010, A&A, 516, A66
Rennehan D., Babul A., Hopkins P. F., Davé R., Moa B., 2019, MNRAS, 483, 3810
Rezaeikh S., Javadi A., Khosroshahi H., van Loon J. T., 2014, MNRAS, 445, 2214
Rice T. S., Goodman A. A., Bergin E. A., Beaumont C., Dame T. M., 2016, ApJ, 822, 52
Riquelme M. A., Spitkovsky A., 2009, ApJ, 694, 626
Rubele S. et al., 2015, MNRAS, 449, 639
Rupke D., 2018, Galaxies, 6, 138
Ruszkowski M., Yang H.-Y. K., Zweibel E., 2017, ApJ, 834, 208
Salem M., Bryan G. L., 2014, MNRAS, 437, 3312
Salem M., Bryan G. L., Corlies L., 2016, MNRAS, 456, 582
Schmidt W., Federrath C., Klessen R., 2008, Phys. Rev. Lett., 101, 194505
Shtykovskiy P., Gilfanov M., 2005, MNRAS, 362, 879
Simpson C. M., Pakmor R., Marinacci F., Pfrommer C., Springel V., Glover S. C. O., Clark P. C., Smith R. J., 2016, ApJ, 827, L29
Skilling J., 1971, ApJ, 170, 265
Skilling J., 1975, MNRAS, 172, 557
Smagorinsky J., 1963, Mon. Weather Rev., 91, 99
Snodin A. P., Shukurov A., Sarson G. R., Bushby P. J., Rodrigues L. F. S., 2016, MNRAS, 457, 3975
Socrates A., Davis S. W., Ramirez-Ruiz E., 2008, ApJ, 687, 202
Spitzer L., Härm R., 1953, Phys. Rev., 89, 977
Squire J., Hopkins P. F., 2018a, MNRAS, 477, 5011
Squire J., Hopkins P. F., 2018b, ApJ, 856, L15
Sridhar S., Goldreich P., 1994, ApJ, 432, 612
Strong A. W., Moskalenko I. V., 2001, Adv. Space Res., 27, 717
Strong A. W., Moskalenko I. V., Ptuskin V. S., 2007, Ann. Rev. Nucl. Part. Sci., 57, 285
Su K.-Y., Hopkins P. F., Hayward C. C., Faucher-Giguère C.-A., Kereš D., Ma X., Robles V. H., 2017, MNRAS, 471, 144
Su K.-Y., Hayward C. C., Hopkins P. F., Quataert E., Faucher-Giguère C.-A., Kereš D., 2018, MNRAS, 473, L111
Su K.-Y. et al., 2019, MNRAS, 487, 4393
Su K.-Y. et al., 2020, MNRAS, 491, 1190
Tang Q.-W., Wang X.-Y., Tam P.-H. T., 2014, ApJ, 794, 26
Thomas T., Pfrommer C., 2019, MNRAS, 485, 2977
Trotta R., Jóhannesson G., Moskalenko I. V., Porter T. A., Ruiz de Austri R., Strong A. W., 2011, ApJ, 729, 106
Tumlinson J., Peeples M. S., Werk J. K., 2017, ARA&A, 55, 389
Uhlig M., Pfrommer C., Sharma M., Nath B. B., Enßlin T. A., Springel V., 2012, MNRAS, 423, 2374
van de Voort F., Quataert E., Hopkins P. F., Faucher-Giguère C.-A., Feldmann R., Kereš D., Chan T. K., Hafen Z., 2016, MNRAS, 463, 4533
van Marle A. J., Casse F., Marcowith A., 2019, MNRAS, 490, 1156
Vladimirov A. E., Jóhannesson G., Moskalenko I. V., Porter T. A., 2012, ApJ, 752, 68
Voelk H. J., 1975, Rev. Geophys. Space Phys., 13, 547
Völk H. J., 1973, Ap&SS, 25, 471
Volk H. J., McKenzie J. F., 1981, Proc. Sci. ICRC, 9, 246
Wang X., Fields B. D., 2018, MNRAS, 474, 4073
Webber W. R., 1998, ApJ, 506, 329
Weisz D. R., Dolphin A. E., Skillman E. D., Holtzman J., Dalcanton J. J., Cole A. A., Neary K., 2013, MNRAS, 431, 364
Wentzel D. G., 1968, ApJ, 152, 987
Wheeler C. et al., 2017, MNRAS, 465, 2420
Wiener J., Oh S. P., Guo F., 2013a, MNRAS, 434, 2209
Wiener J., Zweibel E. G., Oh S. P., 2013b, ApJ, 767, 87
Wiener J., Pfrommer C., Oh S. P., 2017, MNRAS, 467, 906
Wilke K., Klaas U., Lemke D., Mattila K., Stickel M., Haas M., 2004, A&A, 414, 69
Wojaczyński R., Niedźwiecki A., 2017, ApJ, 849, 97
Yan H., Lazarian A., 2002, Phys. Rev. Lett., 89, 281102
Yan H., Lazarian A., 2004, ApJ, 614, 757
Yan H., Lazarian A., 2008, ApJ, 673, 942
Yuan Q., Lin S.-J., Fang K., Bi X.-J., 2017, Phys. Rev. D, 95, 083007
Zank G. P., 2014, Lecture Notes in Physics, Vol. 877, Transport Processes in Space Physics and Astrophysics. Springer Science+Business Media, New York
Zirakashvili V. N., Ptuskin V. S., Völk H. J., 2008, ApJ, 678, 255
Zweibel E. G., 2013, Phys. Plasmas, 20, 055501
Zweibel E. G., 2017, Phys. Plasmas, 24, 055402

APPENDIX A: DEFAULT DAMPING RATES OF GYRO-RESONANT ALFVÉN WAVES

In SC models (Section 3.3), the damping rate Γ of gyro-resonant Alfvén waves ($\delta\boldsymbol{B}[r_{\rm L}]$ or e_A) plays a central role. In the ISM/CGM, it is generally assumed that Γ is dominated by a combination of ion-neutral ($\Gamma_{\rm in}$), turbulent ($\Gamma_{\rm turb}$), linear Landau ($\Gamma_{\rm LL}$), and non-linear Landau ($\Gamma_{\rm NLL}$) damping. Zweibel (2017) and Thomas & Pfrommer (2019) summarize literature estimates of these damping rates from quasi-linear theory, which we adopt as our 'default' set of damping rates, reviewed below.

(i) **Ion-neutral damping:** This is well defined for a partially neutral, hydrogen–helium plasma, giving:[21]

$$\Gamma_{\rm in} = \frac{\alpha_{\rm iH} + \alpha_{\rm iHe}}{2\,\rho_{\rm i}} \sim 10^{-9}\,{\rm s}^{-1}\,f_{\rm neutral}\,T_{1000}^{1/2}\,\rho_{-24}. \tag{A1}$$

Here, $\rho_{\rm i}$ is the mass density of ions, $\alpha_{\rm iX} \equiv (4/3)\,n_{\rm i}\,n_{\rm X}\,\sigma_{\rm iX}\,\sqrt{8\,m_{\rm iX}\,k_{\rm B}T/\pi}$ where ${\rm X} \in \{{\rm H},\,{\rm He}\}$, $m_{\rm iX} \equiv m_{\rm i}\,m_{\rm X}/(m_{\rm i}+m_{\rm X})$, $m_{\rm i}$ and $m_{\rm X}$ are the ion and species X masses (and $n_{\rm i}$, $n_{\rm X}$ their number densities), $\sigma_{\rm iH} = 10^{-14}\,{\rm cm}^2$,

[21] In the neutral ISM at the densities we resolve in our simulations (e.g. GMCs), we can just treat the hydrogen and helium terms here and safely neglect metal ions and charged dust in equation (A1).

and $\sigma_{\rm i\,He} = 3\times10^{-15}\,{\rm cm}^2$, and the latter expression assumes an H mass fraction $\approx$0.75 and defines $T_{1000}\equiv T/1000\,{\rm K}$, $\rho_{-24}\equiv\rho/10^{-24}\,{\rm g\,cm^{-3}}$, and neutral fraction $f_{\rm neutral}=(1-f_{\rm ion})$.

(ii) **Turbulent damping:** Non-resonant motions will interact with and shear gyro-resonant Alfvén waves: accurately capturing this requires understanding the non-linear behaviour of turbulence on scales $\sim r_{\rm L}$, so it remains highly uncertain. Most estimates follow Farmer & Goldreich (2004), and assume a Goldreich & Sridhar (1995) spectrum for 'strong' Alfvénic turbulence with an Alfvén Mach number $\mathcal{M}_A[\ell_A]\equiv|\delta\boldsymbol{v}[\ell_A]|/v_A^{\rm ideal}=1$ at a scale $\ell_A\approx\mathcal{M}_A^{-3}\,\ell_{\rm turb}$, giving $\Gamma_{\rm turb}\sim v_A^{\rm ideal}\,(k_{\rm L}\,k_{\rm turb,\,A})^{1/2}$. Here, $k_{\rm L}\sim1/r_{\rm L}$ and $k_{\rm turb,\,A}\sim1/\ell_A$ represent the resonant and injection wavenumbers, and stand in for appropriate averages over direction and wavenumber (meaning there is order-unity ambiguity here), giving

$$\Gamma_{\rm turb}\equiv\frac{v_A^{\rm ideal}}{r_{\rm L}^{1/2}\,\ell_A^{1/2}}\,f_{\rm cas}\sim2\times10^{-11}\,{\rm s^{-1}}\,\delta v_{\rm turb,\,10}^{3/2}\,\ell_{\rm turb,\,kpc}^{-1/2}\,\rho_{-24}^{1/4}\,\gamma_{\rm L}^{-1/2}\,f_{\rm cas},\tag{A2}$$

where $\delta v_{\rm turb,\,10}\equiv|\delta\mathbf{v}_{\rm turb}[\ell_{\rm turb}]|/10\,{\rm km\,s^{-1}}$ and, as in Section 3.2, we represent our ignorance of the details of turbulence with the parameter $f_{\rm cas}$ (discussed in Section 3.3.5).

(iii) **Linear Landau damping:** This is closely related to turbulent damping, and represents damping of oblique waves whose electric fields interact with the gas via Landau resonance when the propagation angle of the Alfvén waves relative to the local magnetic field is changing owing to turbulent motions (Zweibel 2017). As a result, $\Gamma_{\rm LL}\approx(\pi^{1/2}/4)\,c_s/(r_{\rm L}^{1/2}\,\ell_A^{1/2})\,f_{\rm cas}$ scales with the local turbulent cascade time in exactly the same manner as $\Gamma_{\rm turb}$, but with a different pre-factor. So following Zweibel (2017), we can write

$$\Gamma_{\rm LL}\approx\frac{\sqrt{\pi}}{4}\,\frac{c_s}{v_A^{\rm ideal}}\,\Gamma_{\rm turb}\sim0.4\,\beta^{1/2}\,\Gamma_{\rm turb}.\tag{A3}$$

(iv) **Non-linear Landau damping:** This represents wave-wave interactions, scaling non-linearly with the Alfvén-wave energy $e_{A\pm}$. For a given $e_{A\pm}$, $\Gamma_{\rm NLL,\pm}\approx(e_{A\pm}/e_{\rm B})\,\sqrt{\pi}\,c_s\,k_{\rm L}/8$ (Volk & McKenzie 1981). As shown in Appendix B below, if we assume local quasi-steady-state equilibrium of the Alfvén energy and CR transport coefficients, we do not need to explicitly evolve the $e_{A\pm}$ terms but obtain the 'effective' non-linear damping rate $\langle\Gamma_{\rm NLL}\rangle\approx\Gamma_{\rm NLL}(\langle e_{A\pm}\rangle)$, which becomes

$$\langle\Gamma_{\rm NLL}\rangle\equiv\left[\frac{(\gamma_{\rm cr}-1)\,\pi^{1/2}}{8}\left(\frac{c_s\,v_A}{r_{\rm L}\,\ell_{\rm cr}}\right)\left(\frac{e_{\rm cr}}{e_{\rm B}}\right)\right]^{1/2}\sim0.7\times10^{-11}\,{\rm s^{-1}}\left(\frac{e_{\rm cr,\,eV}}{\gamma_{\rm L}\,\ell_{\rm cr,\,kpc}}\right)^{1/2}\left(\frac{T_{10000}}{f_{\rm ion}\,\rho_{-24}}\right)^{1/4}.\tag{A4}$$

APPENDIX B: NON-EQUILIBRIUM MODEL AND DERIVATION OF THE LOCAL, QUASI-STEADY CR TRANSPORT PARAMETERS

B1 Non-equilibrium scattering rate expressions

Begin from the non-equilibrium CR flux and gyro-resonant Alfvén-wave dynamics equations as derived in Thomas & Pfrommer (2019). Their expression for $e_{\rm cr}$ is identical to ours (see Paper I), with the definition $\Lambda_{\rm st}\rightarrow\mathbf{v}_A\cdot(\boldsymbol{g}_+-\boldsymbol{g}_-)$, where the $\boldsymbol{g}_\pm\equiv(\gamma_{\rm cr}-1)\,(\boldsymbol{F}\mp\boldsymbol{v}_A\,h_{\rm cr})/\kappa_\pm$ and associated $e_{A\pm}\approx|\delta\boldsymbol{B}[r_{\rm L}]|^2/4\pi$ represent the scattering rates and energy in un-resolved Alfvén waves propagating in the $\pm\hat{\boldsymbol{b}}$ directions. Their expressions for the CR flux $\boldsymbol{F}$ and $e_{A\pm}$ are then

$$\frac{\mathbb{D}_t\boldsymbol{F}}{c^2}+\nabla_\parallel P_{\rm cr}=-(\boldsymbol{g}_++\boldsymbol{g}_-),\tag{B1}$$

$$\frac{\partial e_{A\pm}}{\partial t}+\nabla\cdot[\boldsymbol{u}\,h_{A\pm}\pm\boldsymbol{v}_A\,e_{A\pm}]=\boldsymbol{u}\cdot P_{A\pm}\pm\boldsymbol{v}_A\cdot\boldsymbol{g}_\pm-\Gamma_\pm\,e_{A\pm},\tag{B2}$$

where $h_{A\pm}\equiv e_{A\pm}+P_{A\pm}$, $P_{A\pm}\equiv e_{A\pm}/2$, and $\Gamma_\pm$ includes all the damping terms in Appendix A. In the gas momentum equation ($\partial\rho\,\boldsymbol{u}/\partial t$), we explicitly add $P_{A+}+P_{A-}$ to the total (magnetic+thermal+CR) pressure, and the additional 'source' term $\nabla_\parallel P_{\rm cr}+\boldsymbol{g}_++\boldsymbol{g}_-=c^{-2}\mathbb{D}_t\boldsymbol{F}$, to ensure manifest momentum conservation. The damped Alfvén-wave energy ($\Gamma_+\,e_{A+}+\Gamma_-\,e_{A-}$) is added to the gas thermal energy equation (i.e. it is converted from the explicitly tracked Alfvén-energy to thermal energy) instead of directly adding the 'streaming losses' to the thermal energy. The system is closed by the relation:

$$\frac{c\,r_{\rm L}}{\kappa_\pm}=\frac{9\pi}{16}\left(\frac{e_{A\pm}}{e_{\rm B}}\right)\left(1+\frac{2\,v_A^2}{c^2}\right).\tag{B3}$$

With these changes, our equations for the gas momentum and energy, CR energy and flux, and Alfvén-wave energy are exactly identical to the system of equations in Thomas & Pfrommer (2019).

B2 Local equilibrium expressions

Now assume that the CR flux and Alfvén energy equations have reached local steady state ($\partial/\partial t\rightarrow0$, $\mathbb{D}_t\rightarrow0$), and the advection terms (usually smaller by $\sim\mathcal{O}(|\boldsymbol{u}|/c)$ compared to other terms) are negligible. In $e_{A\pm}$, one of the $\pm$ terms – specifically the one corresponding to waves propagating down the CR pressure gradient (i.e. with the same sign along $\pm\hat{\mathbf{b}}$) to the direction of $-\nabla_\parallel P_{\rm cr}$ – will have its corresponding $\pm\mathbf{v}_A\cdot\boldsymbol{g}_\pm$ term be positive-definite, competing against damping, while the other is purely damped. Thus, the antiparallel $e_{A\pm}\rightarrow0$, which implies the corresponding $\boldsymbol{g}_\pm\propto1/\kappa_\pm\propto e_{A\pm}\rightarrow0$ as well. Let us denote the 'surviving' $e_{A\pm}\rightarrow e_A$ and $\boldsymbol{g}_\pm\rightarrow\boldsymbol{g}$. Note that if we write $\boldsymbol{g}\equiv(\gamma_{\rm cr}-1)\,(\boldsymbol{F}-\mathbf{v}_{\rm st}\,h_{\rm cr})/\kappa_\parallel$, where $\kappa_\parallel$ corresponds to the appropriate 'surviving' $\kappa_\pm$ and $\mathbf{v}_{\rm st}\equiv-v_A\,\nabla_\parallel P_{\rm cr}/|\nabla_\parallel P_{\rm cr}|$, the correct 'sign' of the surviving $\boldsymbol{g}_\pm$ is ensured. So with these definitions in steady state, equation (B1) becomes $\nabla_\parallel P_{\rm cr}=-\boldsymbol{g}$ and the non-vanishing $e_{A\pm}$ equation (equation B2) becomes $0=\pm\boldsymbol{v}_A\cdot\boldsymbol{g}-\Gamma\,e_A$, with $\Lambda_{\rm st}\rightarrow\pm\mathbf{v}_A\cdot\boldsymbol{g}$. Here, the $\pm\mathbf{v}_A$ sign corresponds again to the 'surviving' direction so we can replace $\pm\mathbf{v}_A\rightarrow\mathbf{v}_{\rm st}$, giving $\boldsymbol{g}=(\gamma_{\rm cr}-1)\,(\boldsymbol{F}-\mathbf{v}_{\rm st}\,h_{\rm cr})/\kappa_\parallel=-\nabla_\parallel P_{\rm cr}$ and $\Lambda_{\rm st}=\mathbf{v}_{\rm st}\cdot\boldsymbol{g}=-\mathbf{v}_{\rm st}\cdot\nabla_\parallel P_{\rm cr}=\Gamma\,e_A$.

Note now that $\Lambda_{\rm st}=-\mathbf{v}_{\rm st}\cdot\nabla_\parallel P_{\rm cr}$ has exactly the same form as in our 'default' implementation, and the thermal heating term $\Gamma_+\,e_{A+}+\Gamma_-\,e_{A-}\rightarrow\Gamma\,e_A=\Lambda_{\rm st}$ from damping the un-resolved Alfvén waves is exactly the 'streaming loss' term (i.e. the streaming losses can be added directly to the thermal energy, as we do by default). The added term in the gas momentum equation vanishes: $\nabla_\parallel P_{\rm cr}+\boldsymbol{g}_++\boldsymbol{g}_-\rightarrow\nabla_\parallel P_{\rm cr}+\boldsymbol{g}=\mathbf{0}$. From $\boldsymbol{g}=-\nabla_\parallel P_{\rm cr}$ we also have $\boldsymbol{F}=\kappa_\parallel\,\nabla_\parallel e_{\rm cr}+\mathbf{v}_{\rm st}\,h_{\rm cr}$, i.e. our usual streaming+diffusion approximation with streaming speed $v_{\rm st}=v_A$ and diffusivity $\kappa_\parallel=\kappa_\pm(e_A)$. Because $\Gamma\,e_A=-\mathbf{v}_{\rm st}\cdot\nabla_\parallel P_{\rm cr}$, we can solve for e_A and therefore $\kappa_\parallel$: but we should note that if the damping is non-linear, Γ is itself a function of e_A. For the assumptions in Appendix A, we can write $\Gamma=\Gamma_1+\Gamma_2\,(e_A/e_{\rm B})$, where $\Gamma_1=\Gamma_{\rm in}+\Gamma_{\rm turb}+\Gamma_{\rm LL}$ includes the terms independent of e_A and $\Gamma_2\,(e_A/e_{\rm B})=\Gamma_{\rm NLL}$ gives the next-order terms, and we obtain

$$e_A\rightarrow\langle e_A\rangle\equiv\frac{v_A\,|\nabla_\parallel P_{\rm cr}|}{\Gamma_{\rm eff}}=\frac{(\gamma_{\rm cr}-1)\,v_A}{\ell_{\rm cr}\,\Gamma_{\rm eff}}\,e_{\rm cr},\tag{B4}$$

$$\frac{\kappa_\parallel}{c\,r_{\rm L}}\rightarrow\frac{16}{3\pi}\left(\frac{\ell_{\rm cr}\,\Gamma_{\rm eff}}{v_A}\right)\left(\frac{e_{\rm B}}{e_{\rm cr}}\right),\quad v_{\rm st}\rightarrow v_A\tag{B5}$$

$$\Gamma_{\rm eff} \equiv \Gamma(e_A \rightarrow \langle e_A \rangle) = \Gamma_1 \left(\frac{\psi}{2\,[\sqrt{1+\psi}-1]} \right)$$
$$\approx \Gamma_1 + \Gamma_1 \frac{\psi^{1/2}}{2} \equiv \Gamma_{\rm in} + \Gamma_{\rm turb} + \Gamma_{\rm LL} + \Gamma_{\rm other} + \langle \Gamma_{\rm NLL} \rangle, \quad \text{(B6)}$$

where $\psi \equiv 4\,v_A\,|\nabla_\parallel P_{\rm cr}|\,\Gamma_2/(e_{\rm B}\,\Gamma_1^2)$, and the $\approx$ expression for $\Gamma_{\rm eff}$ is exact in both small and large-ψ limits with $\langle \Gamma_{\rm NLL} \rangle = \Gamma_{\rm NLL}(\langle e_A \rangle) = \Gamma_1\,\psi^{1/2}/2 = (v_A\,|\nabla_\parallel P_{\rm cr}|\,\Gamma_2/e_{\rm B})^{1/2}$ (inserting $\Gamma_2 = \sqrt{\pi}\,c_s\,k_{\rm L}/8$ gives $\langle \Gamma_{\rm NLL} \rangle$ in equation A4).[22]

Finally, using the fact that we can trivially re-write streaming+diffusion as 'pure diffusion' or 'pure-streaming' (Section 2.3), it is convenient to re-write this in 'pure-streaming' form, with $\kappa_\parallel \rightarrow 0$ $v_{\rm st} \rightarrow \bar{v}_{\rm st} = v_A + \kappa_\parallel/(\gamma_{\rm cr}\,\ell_{\rm cr})$, i.e.

$$\bar{v}_{\rm st} \rightarrow v_A \left[1 + \frac{4\,c\,r_{\rm L}\,\Gamma_{\rm eff}}{\pi\,v_A^2} \left(\frac{e_{\rm B}}{e_{\rm cr}} \right) \right]. \quad \text{(B7)}$$

Thus, we see that in local steady state, the full Thomas & Pfrommer (2019) expressions reduce to our default expressions with the appropriate $v_{\rm st} = v_A$ and $\kappa_\parallel$. Because, in steady state, $e_A \ll e_{\rm B}$ is miniscule, the non-linear effects of heating and/or pressure changes as the gyro-resonant Alfvén-wave distribution reaches this equilibrium are negligible. And the time-scale to reach this equilibrium is rapid: equation (B2) approaches local equilibrium on the damping time-scale $\sim \Gamma^{-1} \sim 3000\,{\rm yr}$ in the warm ISM and $\sim 30\,{\rm yr}$ in the neutral ISM, while equation (B1) should approach steady state on the scattering time-scale $\sim \kappa/c^2 \sim 10\,{\rm yr}$ (for $\kappa \sim 3\times10^{29}\,{\rm cm^2\,s^{-1}}$).

B3 Behaviour of solutions: neither streaming nor diffusion

Despite the language above, there are three important ways in which the solutions to the CR energy equation (equation 1) for SC models differ from either a traditional streaming equation ($\boldsymbol{F} = \boldsymbol{v}_{\rm st}\,h_{\rm cr}$, with $\mathbf{v}_{\rm st}$ constant) or traditional diffusion equation ($\boldsymbol{F} = -\kappa_\parallel\,\nabla_\parallel e_{\rm cr}$, with $\kappa_\parallel$ constant), as often modelled.

First, and probably most important as our main focus in this paper (Section 2.3), $\kappa_\parallel$, $\mathbf{v}_{\rm st}$, and the 'parallel' direction $\hat{\boldsymbol{b}}$ are variable in both space and time. This means an infinite variety of solutions are possible, which need not have any resemblance to the solutions for constant streaming/diffusion models except in an infinitesimally small 'patch' over an infinitesimally small time.

Secondly, if the flux is not in equilibrium ($\mathbb{D}_t \boldsymbol{F} \neq 0$ in equation 2), then obviously equation (1) will not match the expressions for a pure streaming/diffusion equation even if $\mathbf{v}_{\rm st}$ and $\kappa_\parallel$ are constants. Illustrations of this non-equilibrium behaviour for finite c are shown in Jiang & Oh (2018), figs 1, 10 and 15; Thomas & Pfrommer (2019), figs 5 and 6; and Chan et al. (2019), figs B1, B4, and B5.

Thirdly, even if we assume $\mathbb{D}_t \boldsymbol{F} = 0$, that $\mathbf{v}_{\rm st} = \mathbf{v}_A$ has constant magnitude and direction (and $\hat{\boldsymbol{b}}$ does not change), neglect all collisional losses and source injection, and assume the gas has constant $\boldsymbol{u}$, then equation (1) becomes $d_t e_{\rm cr} = \pm\nabla_\parallel F_\Gamma$, where $d_t e_{cr} = \partial e_{\rm cr}/\partial t + \nabla\cdot[(\boldsymbol{u}+\boldsymbol{v}_A)e_{\rm cr}]$ represents simple advection of the CRs with the Alfvén speed relative to the gas, $F_\Gamma \equiv (4\,c\,r_{\rm L}\,e_{\rm B}/\pi\,v_A)\,\Gamma_{\rm eff} \approx (10^5\,{\rm erg\,s^{-1}\,cm^{-2}})\,n_1^{1/2}\,\Gamma_{-11}$ depends only on the gas density and damping rate, and the $\pm$ sign reflects the sign of $\hat{\boldsymbol{b}}\cdot\nabla_\parallel P_{\rm cr}/|\nabla_\parallel P_{\rm cr}|$. But as others have noted, if $\Gamma \neq 0$, this F_Γ term behaves *neither* as a traditional advection/streaming or as a diffusion term. We illustrate this explicitly with a simplified one-dimensional toy problem in Fig. B1. If the linear Γ terms (e.g. ion-neutral, turbulent, linear-Landau) dominate, then Γ and F_Γ are totally independent of the CR properties (though they depend in a complicated manner on gas properties). So F_Γ behaves as a 'source term' which ensures the total flux down the CR pressure gradient matches the 'bottleneck' value set by SC. This behaviour is qualitatively distinct from e.g. a simple variable or super-Alfvénic advection velocity, which would introduce a term $d_t e_{\rm cr} = -\nabla\cdot(\boldsymbol{v}_{\rm advect}\,e_{\rm cr})$, proportional to the CR energy density. If non-linear Landau damping dominates, $F_\Gamma \propto \sqrt{|\nabla_\parallel e_{\rm cr}|}$, with a coefficient dependent on gas but not CR properties. This gives a 'diffusive' flux proportional to $|\nabla_\parallel e_{\rm cr}|^{1/2}$, instead of $\nabla_\parallel e_{\rm cr}$, which again produces qualitatively different behaviour from a standard diffusion equation, with weaker diffusion in the core and superdiffusive 'tails'.

[22] From equation (B4), we can also confirm that the contribution of the gyro-resonant Alfvén waves to the *total* magnetic pressure is vanishingly small, $P_A/P_{\rm B} \rightarrow (8/9\pi)\,(c\,r_{\rm L}/\kappa_\parallel) \sim 3\times10^{-8}\,B_{\mu{\rm G}}^{-1}\,(10^{30}\,{\rm cm^2\,s^{-1}}/\kappa_\parallel)$, so whether or not we separately include $P_{A\pm}$ in the total MHD pressure or fold it into $P_{\rm B} = |\boldsymbol{B}|^2/8\pi$ as in our 'default' models makes no difference.

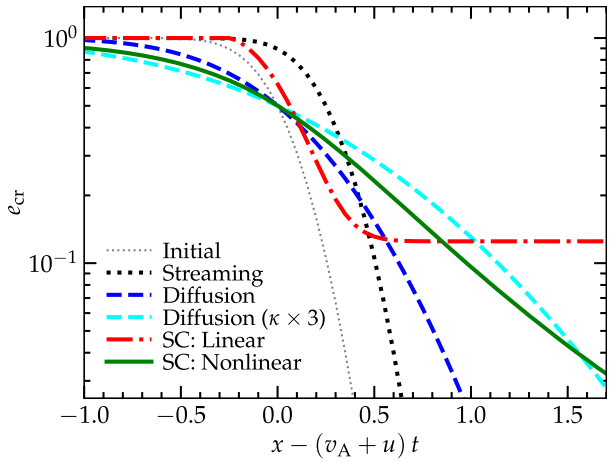

Figure B1. Illustration of the behaviour of the solutions for CR transport with $\kappa_\parallel$ and $v_{\rm st}$ given by SC models (see Appendix B3). We evolve a one-dimensional toy model with parallel fields ($\hat{\boldsymbol{b}} = \hat{\boldsymbol{x}} = \hat{\nabla}_\parallel e_{\rm cr}$) and constant v_A, u, and other background properties. Taking $\tilde{c} \rightarrow \infty$ in equation (2), so $\boldsymbol{F} = F\,\hat{\boldsymbol{x}}$ has its local-equilibrium value, and neglecting sources and sinks, the CR transport equations reduce to $(\partial_t + [v_A + u]\,\partial_x)\,e_{\rm cr} = -\partial_x F$. We consider an initial step-function-like $e_{\rm cr} = 0.5\,{\rm erfc}(x/0.3)$ evolved to time $t = 0.5$ (arbitrary units) assuming (1) traditional streaming/advection with $F = v_{\rm adv}\,e_{\rm cr}$ where $v_{\rm adv} = 1$ is a constant; (2) traditional diffusion with $F = -\kappa\,\partial_x\,e_{\rm cr}$ and $\kappa = 1$ or $=3$ is constant; (3) the expression for $F = F_\Gamma = \kappa_\parallel\,\partial_x e_{\rm cr} = (4\,c\,r_{\rm L}\,e_{\rm B}/\pi\,v_A)\,\Gamma_{\rm eff}$ actually given by SC models (equation B5) assuming linear-damping terms dominate so $\Gamma_{\rm eff} = \Gamma_{\rm in} + \Gamma_{\rm turb} + \Gamma_{\rm LL}$, giving $\partial_x F = -C_{\rm L}\,{\rm SIGN}(\partial_x e_{\rm cr})$ with $C_{\rm L} = 1/2$; (4) the expression for SC (equation B5) with non-linear terms dominant ($\Gamma_{\rm eff} = \Gamma_{\rm NLL}$), so $F_\Gamma = C_{\rm NL}|\partial_x\,e_{\rm cr}|^{1/2}$ with $C_{\rm NL} = 1$. These are the simplest expressions that produce non-trivial behaviour for each version of the equations, and we choose $v_{\rm adv}$, κ, $C_{\rm L}$, $C_{\rm NL}$ so that the 'effective' transport speed $\bar{v}_{\rm st,\,eff}$ is the same around $(x, t) = (0, 0)$. Despite the fact that we can write the SC scalings as a 'diffusion' $\kappa_\parallel$ (equation 7) or 'super-Alfvénic streaming' $\bar{v}_{\rm st}$ (equation 9), the behaviour of even the simplest solutions is not the same as true diffusion or streaming/advection equations.

APPENDIX C: DIFFUSION COEFFICIENTS FOR FAST-MODE SCATTERING

Here, we briefly summarize the scattering rate via fast modes we adopt, directly following the assumptions in Yan & Lazarian (2004), Yan & Lazarian (2008) [YL04]. Begin with the usual expressions for the κ as a function of the pitch-angle diffusion coefficient $D_{\mu\mu}$ for relativistic CRs ($|\mathbf{v}_{\rm cr}| \approx c$), where $\mu = \cos\theta_p$ for pitch angle

θ_p: $\kappa_\parallel = c\,\lambda_{\rm mfp}/3 = (c^2/4)\int_0^1 {\rm d}\mu\,(1-\mu^2)^2\,D_{\mu\mu}^{-1}$. Then define the mode angle $\xi \equiv |\cos\theta_k| = |\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{b}}| = k_\parallel/k$ (with $k_\perp \equiv (1-\xi^2)^{1/2}\,k$), driving scale $\ell_{\rm turb}$, dimensionless wavenumber $\tilde{k} \equiv k\,\ell_{\rm turb}$ and $\tilde{r} \equiv r_L/\ell_{\rm turb}$, and large-scale $|\boldsymbol{B}| = B_0$. YL04 then adopt the expression from Voelk (1975, equation 45 therein) for $D_{\mu\mu}$, keeping only the $n = 0$ (TTD) and $n = \pm1$ (gyro-resonant) terms, and dropping the Alfvénic terms. They assume that fast modes have an isotropic $k^{-3/2}$ power spectrum with ${\rm d}^3\boldsymbol{k}\,I^M(\mathbf{k}) = \mathcal{M}_A^2\,(B_0^2/8\pi)\,\tilde{k}^{-3/2}\,{\rm d}\tilde{k}\,{\rm d}\xi$ from the driving scale to some damping scale $k_{\rm damp}(\xi)$ that is angle dependent, with zero power outside this range, giving

$$D_{\mu\mu}^{(n)} = \frac{\mathcal{M}_A^2\,\Omega\,(1-\mu^2)}{4\pi}\int_0^1 {\rm d}\xi\int_1^{\tilde{k}_{\rm damp}(\xi)}\frac{\xi^2}{\tilde{k}^{3/2}}\left[J_n'(x)\right]^2 R_n\,{\rm d}\tilde{k}, \quad \text{(C1)}$$

where $J_n'(x) = {\rm d}J_n/{\rm d}x$ is the derivative of the appropriate Bessel function with $x \equiv k_\perp\,v_{{\rm cr},\perp}/\Omega = \tilde{k}\,\tilde{r}\,(1-\xi^2)^{1/2}\,(1-\mu^2)^{1/2}$. YL04 take the 'resonance function' R_n to be $R_n = (\pi^{1/2}/\Delta)\,\exp(-q^2/\Delta^2)$ where $q = (k_\parallel\,v_\parallel - \omega_{\rm fast} \pm n\,\Omega)/\Omega \approx k\,r_L\,\xi\,\mu - n$ and $\Delta \equiv k_\parallel\Delta v_\parallel/\Omega \approx k\,r_L\,\xi\,(1-\mu^2)^{1/2}\,\mathcal{M}_A^{1/2}$, or equivalently $R_n = (\pi^{1/2}/\Delta)\,\exp\left[-(\mu - n\,\mu_0)^2/\Delta\mu^2\right]$ with $\Delta\mu^2 \equiv (1-\mu^2)\,\mathcal{M}_A$ and $\mu_0^{-1} \equiv \tilde{k}\,\tilde{r}\,\xi$, as a result of the YL04 assumption that the resonance is broadened with $\Delta\mu \sim \Delta v_\parallel/v_\perp \sim \langle(|\boldsymbol{B}| - B_0)^2\rangle^{1/4}/B_0^{1/2} \sim \mathcal{M}_A^{1/2}$. Defining $\tilde{D}_{\mu\mu} = (D_{\mu\mu}^0 + D_{\mu\mu}^1)/\Omega$, we have $\kappa_\parallel/(c\,r_L) = (1/4)\int_0^1 {\rm d}\mu\,(1-\mu^2)\,\tilde{D}_{\mu\mu}^{-1}$ and the integrals can now be evaluated numerically given $\mathcal{M}_A$, $\tilde{r}$, and $\tilde{k}_{\rm damp}(\xi)$. We follow YL04 to calculate $k_{\rm damp}$ by assuming this is where the damping time becomes shorter than the cascade time, assuming a $k^{-3/2}$ spectrum with $t_{\rm cas}^{-1} \approx (k/\ell_A)^{1/2}\,v_A$, and setting this equal to $\Gamma_{\rm damp}(k,\,\xi,\,\ldots)$ from the sum of collisionless, anisotropic viscous (Braginskii), ion-neutral, and other damping sources (using the expressions in appendix A of YL04).

The simple expressions quoted in the main text are approximate fits to these numerical results over the dynamic range of interest here. They can be approximately derived as follows. When collisionless damping dominates, *if* parallel fast modes are undamped ($f_{\rm ion} = 1$ and $\beta \ll 1$), then the gyro-resonant term ($n = 1$) is sub-dominant in κ and depends relatively weakly on plasma properties (see YL04 discussion), implying that the scaling for $\kappa_\parallel$ is dominated by the TTD ($n = 0$) term. Ignoring $\mu \to 1$ (where the $n = 1$ term dominates), the broad resonance assumption means $R_0 \sim 1$, and because the rigidity is small $J_n'(x) \approx x/2 \sim \tilde{k}\,\tilde{r}$, and $\Gamma_{\rm damp} \sim (\pi\,\beta\,m_e/16\,m_p)\,k\,v_A\,f(\xi)$ where $f(\xi) \sim 1$ for ξ not too close to 0 or 1. Combining all of the ξ, μ integrals into a dimensionless function $g(\xi,\,\mu,\,\mathcal{M}_A) \sim 1$ we can then extract the dimensional scaling for $\kappa_\parallel \sim (c^2/D_{\mu\mu}^0)\,g(\ldots) \sim c\,\ell_A\,(\lambda_{\rm damp}/\ell)^{1/2}$ with $\lambda_{\rm damp}/\ell_A \sim (\beta\,m_e/m_p)$. When viscous damping dominates (again assuming $f_{\rm ion} = 1$ and $\beta \ll 1$), the resonant $n = 1$ term dominates $\kappa_\parallel$ (at $\gamma_{\rm L} \lesssim 100$). Even with $\Delta\mu \sim 1$, the resonant $\mu_0 \sim 1/k\,r_{\rm L}$ term in R_1 is large unless $k \gtrsim 1/r_{\rm L}$, which for a $\beta \ll 1$ viscous damping rate of $\Gamma_{\rm visc}(\beta < 1) \approx k^2\,\nu_{\rm v}\,(1-\xi^2) \sim 2\,k^2\,\nu_v\,\epsilon_\xi$ (defining $\epsilon_\xi = 1 - \xi$) requires $|\epsilon_\xi| \ll 1$, such that $k_{\rm damp} \gg 1/r_{\rm L}$. Taking these limits and evaluating gives $\kappa_\parallel$ inversely proportional to powers of $\epsilon_\xi \sim \tilde{r}^{3/2}\,(\ell_A\,v_A/\nu_{\rm v})$.

Finally, regardless of what dominates $\Gamma_{\rm damp}$, if the parallel ($\xi \approx \pm1$) modes are damped on scales $k_{\rm damp}(\xi \to 1) \ll 1/r_{\rm L}$, then $R_1 \to 0$ rapidly as $\exp\left[-(k_{\rm damp}\,r_{\rm L})^{-2}\right]$, and as a result $\kappa_\parallel \to \infty$ as we integrate to $\mu \to 1$ (*regardless* of the behaviour of the TTD terms and broadening $\Delta\mu \sim \mathcal{M}_A^{1/2} \sim 1$). This occurs with ion-neutral damping ($\Gamma_{\rm damp} = \Gamma_{\rm in}$, independent of ξ), which gives $k_{\rm damp}\,r_{\rm L} \approx (f_{\rm neutral}/f_{\rm n,\,0})^{-2} \lesssim 1$ where $f_{\rm n,\,0} = 0.001\,(n_1\,\beta)^{-3/4}\,T_4^{1/4}\,(\ell_{\rm turb,\,kpc}\,\gamma_{\rm L})^{-1/2}$. It also occurs if $\beta \geq 1$, in which case the viscous damping becomes strong as $\xi \to 1$ with $\Gamma_{\rm visc} \approx k^2\,\nu_{\rm v}\,|3\,\xi^2 - 1|$, giving $k_{\rm damp}\,r_{\rm L} \ll 10^{-4}$ for any physically plausible parameters with Braginskii ν_v. These give the damping 'cut-offs' used in the text (Section 3.2.1): $f_{\rm cut} = \exp\{(f_{\rm neutral}/f_{\rm n,\,0})^4 + (\beta/0.1)^{1.5}\}$.

APPENDIX D: ADDITIONAL PHYSICAL AND NUMERICAL VARIATIONS EXPLORED

Here and in Papers I and II, we have considered a large number of additional tests to confirm that the dominant uncertainty in CR transport is the form of κ_*, as opposed to e.g. numerical uncertainties or the detailed form of the transport equation. These include the following:

(i) **Equilibrium versus non-equilibrium transport expressions:** This is discussed explicitly in the text (and see Appendix B above), but we list it here for completeness.

(ii) **Maximum 'free-streaming' speeds:** $\tilde{c}$ represents the 'effective speed of light' which determines the maximum free-streaming speed of CRs. In Papers I and II, we show this is a 'nuisance parameter', because the local steady-state CR flux and energy converge to the same values independent of $\tilde{c}$, so long as it is larger than local advection/diffusion speeds. In addition, we have tested all the models in this paper assuming $\tilde{c} = 500\,{\rm km\,s^{-1}}$ or $\tilde{c} = 1000\,{\rm km\,s^{-1}}$ as well as $\tilde{c} = {\rm MAX}(1000\,{\rm km\,s^{-1}},\,2\,\kappa_*/\ell_{\rm cr})$ (our default). So long as $\tilde{c} \gtrsim \kappa_*/\ell_{\rm cr} \sim 300\,{\rm km\,s^{-1}}\,\tilde{\kappa}_{29}/\ell_{\rm cr,\,kpc}$, then the results are robust to $\tilde{c}$; for the highest $\kappa_* \gg 10^{30}\,{\rm cm^2\,s^{-1}}$ runs here, this means we require $\tilde{c} \gtrsim 1000\,{\rm km\,s^{-1}}$ to ensure converged results (otherwise L_γ is artificially large because CRs are 'slowed down'), but even in this limit the qualitative conclusion that CRs escape efficiently is robust.

(iii) **Explicit perpendicular diffusion:** As shown in Papers I and II, even assuming pure isotropic diffusion leads only to a factor ∼2–3 lower κ_* required to reproduce the same observed L_γ, grammage, etc. We confirm this in limited tests of our constant-κ and 'SC100' models. Physically, we generally expect the perpendicular diffusivity to be suppressed by a factor $\sim r_{\rm L}/\lambda_{\rm mfp}$: we have experimented with models that explicitly include perpendicular diffusive flux $F_\perp = \kappa_\perp\,(\nabla - \nabla_\parallel)e_{\rm cr}$ where $\kappa_\perp = (r_{\rm L}/\lambda_{\rm mfp})\,\kappa_\parallel \approx r_{\rm L}\,c/3$ and find (as expected) this makes a negligible difference compared to assuming pure parallel diffusion.

(iv) **Resolution:** We emphasize the importance of resolving the ISM/CGM in the text, yet it is reasonable to worry that the smallest molecular clouds and star-forming regions are under-resolved. Despite this, we have shown in previous papers that GMC properties in these simulations including their size–mass relations (mean densities), linewidth–size relations, mass functions, magnetic field strengths, and lifetimes agree well with observations and appear converged down to clouds with as few as ∼10 resolution elements (Hopkins et al. 2018b; Orr et al. 2018; Grudić et al. 2019; Guszejnov, Hopkins & Graus 2019; Orr, Hayward & Hopkins 2019; Benincasa et al. 2020; Guszejnov et al. 2020; Keating et al. 2020). While this excludes the smallest clouds at our resolution, it includes the complexes that contain > 90 per cent of all galactic star formation (Rice et al. 2016). And as shown in the main text, our key conclusions are not particularly sensitive to the behaviour of CRs in the most dense, neutral ISM because of its small volume-filling fraction. Moreover, Papers I and II consider extensive explicit resolution tests, in both cases varying the mass resolution of the 'constant-κ' models by factors of ∼100. In both cases (consistent with further extensive resolution studies in Hopkins et al. 2018b), we showed that our predictions for dwarfs were only weakly sensitive to resolution. For MW-mass galaxies some galaxy properties do depend on resolution

(e.g. the central regions of the galaxies tend to be more dense at lower resolution, owing to less efficient resolution of galactic outflow 'venting'); however, the qualitative effects of CRs, and range of allowed transport parameters, were robust to resolution. As $\Sigma_{\rm central}$ changed (weakly) with resolution, the corresponding $L_\gamma/L_{\rm sf}$ shifts along the ellipses for a given, single-resolution (i.e. systems move along the relations in Fig. 3, for fixed CR transport parameters). We have confirmed this result in our simulations without a constant κ by running several of the models here (four ET models and four SC models) for each of (**m11i**, **m11f**, **m12i**) at factor ∼8 lower mass resolution (run initially to test and validate our implementation). In more limited tests of **m12i** at $z\sim 0$ we have also confirmed that the exact choices for force softening and star formation criteria have no substantial effects on our conclusions.

(v) **Form of the CR flux time derivative:** The CR flux equation, equation (2), has subtle ambiguities related to the frame in which the CR flux is evaluated, order in $\mathcal{O}(v/c)$, assumptions about the form of the CR distribution function, and extrapolation of scattering terms from quasi-linear theory. These are discussed in e.g. Zweibel (2017), Thomas & Pfrommer (2019), Chan et al. (2019), and references therein, and explored in Papers I and II, but we briefly discuss them here. The formulations of CR transport in Chan et al. (2019), Jiang & Oh (2018), and Thomas & Pfrommer (2019), as well as simpler 'pure diffusion/streaming' models commonly adopted in the literature are – for a *specific* value of the local κ_* (i.e. assuming that $|\delta\boldsymbol{B}[r_{\rm L}]|^2$ has taken on some local quasi-equilibrium value) – identical up to the form of the operator $\mathbb{D}_t\mathbf{F}$ in equation (2). In the 'pure diffusion/streaming' model, $\mathbb{D}_t\boldsymbol{F}=\mathbf{0}$, so $\boldsymbol{F}\equiv-\kappa_*\,\nabla_\parallel e_{\rm cr}$ and there is no flux equation to solve (simply a single advection+diffusion equation for $e_{\rm cr}$). In Paper I, $\mathbb{D}_t\boldsymbol{F}=\partial\boldsymbol{F}/\partial t+\nabla\cdot(\boldsymbol{u}\otimes\boldsymbol{F})$, and in Jiang & Oh (2018) $\mathbb{D}_t\boldsymbol{F}=(\hat{\boldsymbol{F}}\otimes\hat{\boldsymbol{F}})\cdot[\partial(\boldsymbol{F}+\boldsymbol{u}\,h_{\rm cr})/\partial t]$; neither of these papers attempted to derive the flux equation from first principles, but rather simply adopted a form (inspired by two-moment treatments of radiation hydrodynamics and similar problems) which relaxes to the correct behaviour in various limits. Thomas & Pfrommer (2019) do attempt such a derivation, and obtain $\mathbb{D}_t\boldsymbol{F}\equiv\hat{\boldsymbol{F}}\,[\partial|\boldsymbol{F}|/\partial t+\nabla\cdot(\boldsymbol{u}\,|\boldsymbol{F}|)+\boldsymbol{F}\cdot\{(\hat{\boldsymbol{F}}\cdot\nabla)\,\boldsymbol{u}\}]=\partial\boldsymbol{F}/\partial t+\nabla\cdot(\boldsymbol{u}\otimes\boldsymbol{F})+(\boldsymbol{F}\cdot\nabla)\,(\boldsymbol{u}_\parallel-\boldsymbol{u}_\perp)$.[23] But all of these are within the $\mathcal{O}(1/\tilde{c}^2)$ term in equation (2), so they vanish when $\tilde{c}\rightarrow\infty$, *or* when the CR flux reaches local quasi-steady state ($\mathbb{D}_t\boldsymbol{F}\rightarrow 0$), which occurs on the extremely short CR mean free path/time defined in Appendix B. In fact, the variants with $\mathbb{D}_t\mathbf{F}\neq\mathbf{0}$ above differ *only* if $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{b}}$ are non-uniform and time dependent, on spatial/time-scales below the CR mean free path (time) $\sim\kappa/\tilde{c}$ ($\sim\kappa/\tilde{c}^2$), when $\tilde{c}$ is relatively small and the CR flux is out of steady state. But this is exactly the regime where adopting $\tilde{c}<c$ means the CR flux differs from the 'true' physical solution, so none of these can be exact. To the extent that our results are converged with respect to $\tilde{c}$, as demonstrated in Papers I and II, they must also be independent of the choice of $\mathbb{D}_t$ here. Moreover, Paper I considers the much more radical choice $\mathbb{D}_t=\mathbf{0}$, and shows the galaxy results are essentially identical. All our constant-κ models have been re-run with the different variant $\mathbb{D}_t$ forms discussed above in Papers I and II, where we showed this had a negligible effect on the observables predicted here. We have repeated this with a limited study of models 'Fast-YL04' and 'SC100' here, where we find the same result.

(vi) **Form of the scattering terms:** Another ambiguity is whether to represent the scattering term in equation (2) as $\boldsymbol{F}/\kappa_*$ with $\kappa_*\equiv\kappa_\parallel+\gamma_{\rm cr}\,v_{\rm st}\,\ell_{\rm cr}$ (our default), or as $(\boldsymbol{F}-\boldsymbol{v}_{\rm st}\,h_{\rm cr})/\kappa_\parallel$, as in Appendix B. Both are consistent with quasi-linear theory, and become exactly identical when $\tilde{c}\rightarrow\infty$ and/or the flux $\boldsymbol{F}$ reaches local quasi-steady state ($\mathbb{D}_t\boldsymbol{F}$ is small), so again our experiments with different $\mathbb{D}_t\boldsymbol{F}$ and $\tilde{c}$ indicate our conclusions are robust to this choice. And because our 'favoured' models have a drift velocity $|\boldsymbol{F}|/h_{\rm cr}\gg v_A$, this is further minimized (generally contributing <5 per cent corrections, re-running different models for select short periods). Moreover, our 'non-equilibrium' model (Section 3.3.2) adopts the $(\boldsymbol{F}-\boldsymbol{v}_{\rm st}\,h_{\rm cr})/\kappa_\parallel$ form and gives similar results to the equilibrium model with κ_*.

(vii) **Form of the 'streaming loss' term:** The 'streaming loss' term, $\Lambda_{\rm st}$ in equation (1) is well motivated in local steady state, SC models (where it takes the form $\Lambda_{\rm st}\approx v_A\,|\nabla_\parallel P_{\rm cr}|$), as it arises from the damping and thermalization of gyro-resonant Alfvén waves (well below our simulation resolution limits) excited by CR streaming (see Appendix B). It is less clear how it should behave in our ET models or models with sub-Alfvénic streaming. We discuss this and vary the term extensively in our constant-κ models in Papers I and II, considering $\Lambda_{\rm st}={\rm MIN}(v_A,\,v_{\rm st})\,|\nabla_\parallel P_{\rm cr}|$ (our default here), or $\Lambda_{\rm st}=\mathbf{v}_{\rm st}\cdot\nabla P_{\rm cr}$, or $\Lambda_{\rm st}=v_A\,|\nabla_\parallel P_{\rm cr}|$, or $\Lambda_{\rm st}=0$. There we showed this had very small ($\sim$ 10 per cent, at high κ) effects on the observables we predicted. Here, we have repeated these comparisons for a subset of our ET models at $z\sim 0$ (restarting them for a short time) to confirm that this produces nearly negligible perturbations to L_γ. We also find that any model where this $\Lambda_{\rm st}$ term is able to produce large CR losses in the ISM or inner CGM (where it might influence our predictions) is already in the well into the regime where collisional losses dominate inside of the galaxy ISM.

(viii) **Exact momentum-conserving formulation:** In our default formulation, we assume a local strong-coupling approximation so the CRs enter the gas momentum equation via the term $\nabla P_{\rm cr}$. As noted in Appendix B, if we approximate the flux equation in the form described therein or in our second-moment expansion equation (2) (both accurate to $\mathcal{O}(v/c)$), then exactly conserving total momentum accounting for the change in inertia of the CRs themselves would require adding a source term $[\nabla_\parallel P_{\rm cr}+\boldsymbol{g}_++\boldsymbol{g}_-]=(\boldsymbol{F}-\boldsymbol{F}_{\rm eqm})/(3\,\kappa_*)=\mathbb{D}_t\,\boldsymbol{F}/\tilde{c}^2$ to the gas momentum (where $\boldsymbol{F}_{\rm eqm}=-\kappa_*\nabla_\parallel e_{\rm cr}$ is the local steady-state flux). This obviously vanishes as $\tilde{c}\rightarrow\infty$ or $|\mathbb{D}_t\boldsymbol{F}|\rightarrow 0$ so our tests of varying $\tilde{c}$, or taking $\mathbb{D}_t\boldsymbol{F}=\mathbf{0}$ exactly, show that the term should not change our results. We do not include this by default because, as noted in Jiang & Oh (2018) and Paper I, if $\tilde{c}\ll c$, this term is artificially large and the CR contribution to the force will be underestimated compared to a converged solution with respect to $\tilde{c}$ (because the CR flux deviation from equilibrium is artificially modified by $\tilde{c}$).

(ix) **Local turbulent velocity estimator:** Because the local turbulent velocities $\delta v_{\rm turb}$ on a scale (of order our simulation resolution) $\ell_{\rm turb}$ appear in the scalings for both ET and SC (via turbulent damping) CR scattering, we have considered four different local on-the-fly estimators for this quantity. (1) Our default, from Hopkins et al. (2013a), $\delta v_{\rm turb}=\|\nabla\otimes\boldsymbol{v}\|\,\ell_{\rm turb}\equiv(\sum_{ij}|\nabla_j\,v_i\,\ell_{\rm turb}|^2)^{1/2}$ the Frobenius norm (sum over components) of the velocity difference across a resolution element estimated from the (non-slope-limited) velocity gradient with $\ell_{\rm turb}=\Delta x=(m_i/\rho_i)^{1/3}$ the resolution scale. (2) The 'shear corrected' norm (norm of the trace-free diagonalized shear tensor of the velocity field, constructed from $\nabla_j v_i$) times Δx, as defined and commonly used for Smagorinsky (1963) 'subgrid-scale'

[23]Note that the Thomas & Pfrommer (2019) formulation only differs from the Paper I formulation by the term $(\boldsymbol{F}\cdot\nabla)\,(\boldsymbol{u}_\parallel-\boldsymbol{u}_\perp)/\tilde{c}^2=[\boldsymbol{F}\{\hat{\boldsymbol{F}}\cdot[(\hat{\boldsymbol{F}}\cdot\nabla)\boldsymbol{u}]\}-(\boldsymbol{F}\cdot\nabla)\,\boldsymbol{u}]/\tilde{c}^2$. This term (1) incorporates a Lorentz term that manifestly ensures $\hat{\boldsymbol{F}}=\hat{\boldsymbol{b}}$ is preserved and (2) includes the 'pseudo-forces' described by Thomas & Pfrommer (2019) which arise because $\boldsymbol{F}$ is defined in the (non-stationary) fluid frame in which the CR distribution function can be assumed to be gyrotropic.

turbulent diffusion models (see e.g. Colbrook et al. 2017; Escala et al. 2018). (3) The direct dispersion $|\delta v_{\rm turb}|_a^2 = \sum_b |\boldsymbol{v}_b - \boldsymbol{v}_a|^2$ across neighbours in a sphere of volume $\ell_{\rm turb}^3$. (4) The more sophisticated (but computationally expensive) method developed in Rennehan et al. (2019), motivated by detailed turbulence studies, where we smooth the velocity field on multiple scales in multiples of the resolution Δx, calculate the relative power in velocity fluctuations, and derive the associated turbulent $E(k)$ at $k \rightarrow 1/\Delta x$. On top of these variations, we also note that many of the models which involve $\delta v_{\rm turb}$ really use this as a proxy for $\delta \boldsymbol{B}_{\rm turb}$, assuming that at the Alfvén scale ℓ_A, $\delta v_{\rm turb} \approx v_A$ and $\delta \boldsymbol{B}_{\rm turb} \sim |\boldsymbol{B}|$. So we have also re-computed *all* of the relevant scalings using $\delta \boldsymbol{B}_{\rm turb}$ measured directly in the code (with the same four estimators described above), to estimate ℓ_A, and extrapolating the relevant assumed power spectra below this scale. We find that although these eight model variants can produce quite large (order-of-magnitude, in some cases) differences in the specific value of $\delta v_{\rm turb}(\boldsymbol{x}, t)$ estimated at any given point $(\mathbf{x}, t)$ in the ISM, the *statistics* produced by the different estimators are quite similar. A more detailed comparison of these in their own right will be the subject of future work, but relevant for this study, integral quantities like L_γ are ultimately altered at the factor $\lesssim$2 level (comparing all these variations), not enough to alter our conclusions.

(x) **Additional statistics (different galaxies):** Given the very large number of different CR transport models we survey here, we chose to limit our study to three representative galaxies or 'zoom-in regions' **m11i**, **m11f**, **m12i** in Table 2. While this is still an improvement over comparing with a single MW model alone, one might worry that our conclusions could be biased by either limited statistical power or systematic effects owing to e.g. the structure or formation history of the particular galaxies. However, we have re-run most of the 'constant-κ' models with a much larger number of simulations, presented in detail in Paper II (along with some additional zoom-in regions of Local Groups following Garrison-Kimmel et al. 2019): altogether >35 zoom-in regions containing several hundred resolved galaxies ranging in $z = 0$ halo mass between $M_{\rm halo} \sim 10^9$–$10^{13}\,{\rm M}_\odot$ (including specifically 10 'single' MW-mass systems and 4 Local Group pairs each containing an MW and Andromeda-like galaxy). We show there that all our conclusions here regarding statistics of e.g. comparison with $L_\gamma/L_{\rm SF}$ and $e_{\rm cr}$, and the inferred observationally allowed values of κ, are robust. We have also run a subset of the non-constant-κ models here ('Alfvén-C00', 'Fast-YL04', 'SC:Default', and 'SC:100') on an expanded halo sample including haloes (**m10q**, **m11q**, **m11g**, **m12f**) from Paper II, with halo masses $\log(M_{\rm halo}/{\rm M}_\odot) \sim$ (10, 11, 11.5, 12) and stellar masses $\log(M_*/{\rm M}_\odot) \sim$ (6.3, 9.0, 10, 10.8), respectively. Each of haloes (**m11q**, **m11g**, **m12f**) behave broadly similarly to our standard (**m11i**, **m11f**, **m12i**), respectively (galaxies with similar mass) for each specific CR transport model. To the extent that they differ in e.g. $L_\gamma/L_{\rm sf}$ they move (slightly) *along*, not with off of, the relation defined by (**m11i**, **m11f**, **m12i**) in Fig. 3. Halo **m10q** (the least massive) is consistent with the extrapolation of these trends, but falls outside the plotted and observed range (with much lower mass/luminosity/density) in our comparisons. All of this is consistent with our larger statistical study in Paper II.

(xi) **CR injection efficiency:** As discussed in Paper II, if we add additional sources of CRs (e.g. structure formation shocks, AGN) then this will further increase L_γ without increasing $L_{\rm SF}$, requiring larger diffusivities to reproduce observations, but these are almost certainly sub-dominant for CR production compared to SNe in the galaxies of interest. If we change the assumed efficiency of CR production in SNe ($\epsilon_{\rm cr}$), in the calorimetric limit this changes $L_\gamma/L_{\rm SF} \propto \epsilon_{\rm cr}$, so reproducing the observations of the SMC/LMC/M33 with, say, $v_{\rm st} \sim v_A$ (so all galaxies are near-calorimetric) while also matching the observed starburst systems would require factor of $\sim$100 variation in $\epsilon_{\rm cr}$ in SNe *as a function of galaxy properties* (which cannot be primarily metallicity, since this is constant for some observed systems with different $L_\gamma/L_{\rm SF}$). More importantly, changing $\epsilon_{\rm cr}$ does *not* change the median grammage or residence time 'per CR', so reproducing the grammage, residence time, and L_γ observations simultaneously, *or* reproducing the L_γ observations in different galaxies simultaneously with a constant $\epsilon_{\rm cr}$, requires $\epsilon_{\rm cr} \sim 0.1$. We have experimented in Paper II with modest variations $\epsilon_{\rm cr} \sim 0.05$–$0.2$: the range of observations and simulation spread in predictions make it difficult to rule out factor $\sim$2 changes in $\epsilon_{\rm cr}$, but at this level these variations have no qualitative effect on our conclusions.

APPENDIX E: COMPARISON TO LOW-DIFFUSION MODELS IN OTHER COSMOLOGICAL SIMULATIONS

Recently, Pfrommer et al. (2017) and Buck et al. (2020, B20) explored the effects of explicit CR transport models in idealized isolated galaxy and cosmological simulations, similar in spirit to our Papers I and II. These simulations used a different code and numerical method with somewhat lower resolution. They also employ a fundamentally different treatment of the ISM wherein any gas above a density $n > 0.1\,{\rm cm}^{-3}$ is assigned a 'stiff' effective (quasi-adiabatic) equation of state, with an SFR set by calibration to observations, and is assumed to launch galactic winds with a mass-loading and velocity set analytically to reproduce the galaxy mass function following Grand et al. (2017). The scheme is designed for large-volume simulations that do not resolve ISM or outflow phase structure, so we might expect significant differences from our results here.

The authors consider three transport models (1) CR advection only ($\kappa_{29} = 0$, $v_{\rm st} = 0$, with no 'streaming loss' term); (2) diffusion-only with $\kappa_{29} = 0.1$ ($v_{\rm st} = 0$, no 'streaming loss'); and (3) diffusion with 'streaming losses' but without streaming motion ($\kappa_{29} = 0.1$, $v_{\rm st} = 0$, but taking the streaming losses to be $v_A|\nabla_\parallel P_{\rm cr}|$ with $v_A \gtrsim 100\,{\rm km\,s^{-1}}$). These are all akin to a subset of our 'constant-diffusivity' models from Papers I and II, with low κ.

Despite the simulation differences, we find that their conclusions are similar to ours, for similarly low diffusivities: Fig. E1 shows this directly. As the authors state directly in Pfrommer et al. (2017) (see fig. 3 therein), in their MW-like haloes, all their models predict that almost all of the injected CR energy is lost to collisions, and so produce $L_\gamma/L_{\rm SF}$ near the calorimetric limit. Moreove, even at LMC and SMC star formation rates their predicted $\dot{E}_{\rm coll}/\dot{E}_{\rm cr} \sim 0.3$ in their favoured model (i.e. they are always within a factor of $\sim$3 of calorimetric). The cosmological simulations in B20 give a similar result (fig. 14 therein): even for the smallest dwarf galaxies (lowest SFRs) plotted, the predicted L_γ is within a factor $\sim$1.5–3 of the calorimetric limit. The other diagnostics we consider here also give consistent results. For example, their models (1) and (2) predict a CR energy density at the solar circle in MW-like galaxies of $e_{\rm cr}(r \approx 8\,{\rm kpc}) \sim 15$–$20\,{\rm eV\,cm^{-3}}$.[24] Where $\dot{E}_{\rm coll}/\dot{E}_{\rm cr} < 1$, we can

[24]Their model (3) predicts a lower value of $e_{\rm cr}(8\,{\rm kpc})$ only because with streaming losses but no streaming transport (and weak diffusion) and unphysically large $v_A \sim 200\,{\rm km\,s^{-1}}$ in the warm ISM (owing to the artificial ISM 'effective equation of state'), the energy loss time-scale from 'streaming' $\sim 3\,\ell_{\rm cr}/v_A$ (see their figs 10 and 12) in their simulations at ~ 8 kpc is $\sim$10 times shorter than the diffusion time [$\sim$(few kpc)$^2/\kappa_{\rm iso}$] for CRs to reach that radius, so most of the CR energy is lost to 'streaming losses' despite the model not including streaming motion.

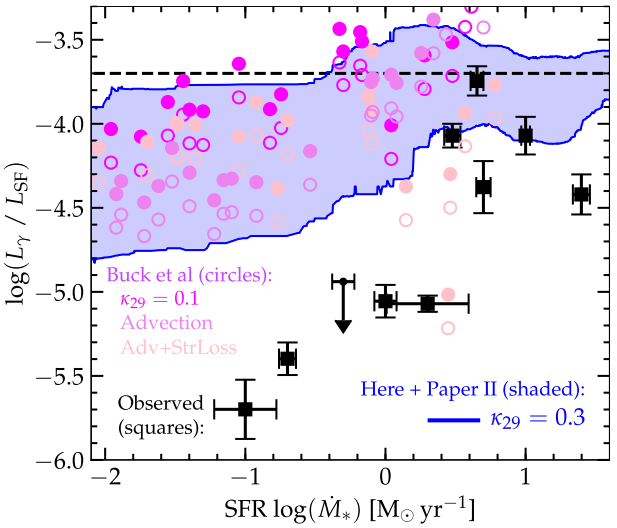

Figure E1. Comparison of γ-ray emission $L_\gamma/L_{\rm SF}$ versus SFR (as Fig. 4) in our low-diffusivity CD model $\kappa_{29} = 0.3$ (shaded shows 2σ range) and observed (black points with error bars). We contrast (see Appendix E) the results from Buck et al. (2020, B20; circles), who predict L_γ from independent cosmological simulations without ISM phase structure, considering low-diffusivity models including (1) $\kappa_{29} = 0.1$ (with $v_{\rm st} = 0$), (2) advection-only ($\kappa_{29} = v_{\rm st} = 0$), (3) advection+streaming losses ($\kappa_{29} = v_{\rm st} = 0$, but still adding a rapid 'streaming loss' sink term $= v_A\,\nabla P_{\rm cr}$ in the CR energy equation). Open points show the values of L_γ and $L_{\rm SF}$/SFR taken exactly as given in B19's fig. 14. Solid points correct these points to adopt the identical stellar and γ-ray bolometric corrections, γ-ray bandpass, and assumptions about hadronic loss rates as those adopted in the text here. Their predictions are nearly identical to ours for similar (low) diffusivity, and predict for $\kappa_{29} \ll 1$ that > 90 per cent of galaxies are within a factor $\sim$3 of the calorimetric limit at any SFR. The B20 models also predict solar-circle grammage $X_s \gtrsim 100\,{\rm g\,cm^{-2}}$, CR energy density $e_{\rm cr} \sim 20\,{\rm eV\,cm^{-3}}$, and residence times $\gg 100$ Myr, similar to our low-κ models in Figs 6 and 7. The predictions for $\kappa_{29} \lesssim 1$ are consistent between simulations and clearly ruled out by both γ-ray and MW observations: per Appendix E, B19's conclusion that low-κ models are observationally permitted stems from not considering MW constraints and from plotting the γ-ray data at the incorrect values of $L_{\rm SF}$.

use their adopted conversion formulae for their predicted γ-ray luminosities and injection rates to directly calculate the grammage in their simulations as well.[25] In all three CR transport models, they predict a grammage in MW-mass systems of $X_s \sim 80$–$200\,{\rm g\,cm^{-2}}$, and for all lower mass/SFR systems (down to $\dot{M}_* \sim 0.001\,{\rm M_\odot\,yr^{-1}}$) they predict $X_s \sim 40$–$130\,{\rm g\,cm^{-2}}$. Finally, although we cannot directly reconstruct their predicted residence times, their predicted $L_\gamma/L_{\rm SF}$ or grammage (given their collisional loss rate and mean ISM densities in B20 fig. 10), or our simple analytic model in Section 5.1.1 all imply similar $\Delta t_{\rm res} \gtrsim 500$ Myr.

Each of these conclusions is similar to those from our similar ($\tilde{\kappa} \leq 0.3$) simulations in Table 1 and Papers I and II. Likely the reason we obtain such good agreement, despite considering very different simulations, is simply because the quantities above 'saturate' once CRs approach the pure-advection/low-diffusion/calorimetric limit. However, B20 claim that their results disagree significantly with ours, arguing that their low-diffusivity models do reproduce the observations. They attribute the difference in predictions primarily to the treatment of dense gas, but as we have shown (1) there is actually very little difference in the predictions and (2) dense gas has little effect on our predictions.

The actual differences stem from how the observations are treated. Pfrommer et al. (2017) and B20 compare only to the L_γ–$\dot{M}_*$ correlation: they do not consider grammage or residence time or CR energy density constraints as we do here (all of which clearly rule out these lower κ models). Moreover, for the L_γ–$\dot{M}_*$ correlation, the authors estimate $\dot{M}_*$ of the observed systems (or, equivalently, the far-IR (FIR) 8–1000 μm luminosity of their simulations) by assuming a universal conversion factor $\dot{M}_*/({\rm M_\odot\,yr^{-1}}) = 1.34\times10^{-10}\,(L_{\rm FIR}/L_\odot)$. However, as noted in both Pfrommer et al. (2017) and B19, it is well known that this correlation and conversion factor break down quite severely in low-SFR systems including the SMC, LMC, and M33 (and even at factor $\sim$2–3 level in the MW and M31), as the conversion they adopt assumes that *all* the light emitted by massive stars is absorbed by cold dust and re-processed into FIR (the particular calibration they adopt is derived for luminous infrared galaxies, with typical extinctions $A_v \sim 100$). For the SMC, this means their adopted SFR ($\sim 0.008\,{\rm M_\odot\,yr^{-1}}$) is a factor $\sim$10–30 lower than implied by high-mass X-ray binary counts (Shtykovskiy & Gilfanov 2005; Haberl & Sturm 2016), young stellar object counts (Hony et al. 2015), long-period variable star counts (Rezaeikh et al. 2014), simple bolometric ultraviolet continuum (Hagen et al. 2017), or Hα emission (Wilke et al. 2004) conversions, or the 'gold standard' (to which many other methods are calibrated) resolved main-sequence turn-offs (i.e. stellar HR or colour–magnitude diagram studies; Harris & Zaritsky 2004; Noël et al. 2009; Indu & Subramaniam 2011; Weisz et al. 2013; Rubele et al. 2015). More importantly, this means their assumed SNe rate (which is what $L_{\rm SF}$ is ultimately used for, to estimate $R_{\rm SNe}$ and therefore $\dot{E}_{\rm cr} \approx 10^{50}\,{\rm erg}\,R_{\rm SNe}$) is $\sim 1/15,000$ yr, a factor $\sim$15–30 lower than inferred from direct observations of SNe remnants in the MCs (Maoz & Badenes 2010; Leahy 2017; Maggi et al. 2019). There are also some differences in the γ-ray spectral slopes/bolometric corrections assumed, as for example B20 include all emission from 0.1 to 100 GeV (likely including non-negligible pulsar contamination), but these are generally smaller (factor $\sim$2) effects.

The net result of this is that the SMC is plotted in e.g. B20 fig. 14 as if it has $L_\gamma \sim 0.4\,L_{\rm calor}$; this, in turn, means that their theoretical predictions with low $\tilde{\kappa}$ appear consistent – as would indeed our own low-diffusivity $\kappa_{29} = 0.3$ model shown in our Fig. 3. However, observational studies of these systems which carefully account for SNe rates and/or UV luminosities and γ-ray spectra place the SMC at $L_\gamma \sim 0.007\,L_{\rm calor}$ (Lacki et al. 2011; Lopez et al. 2018), a factor of $\sim$50 lower. If we compare the grammage, residence time, and/or CR energy density constraints in the MW (see values above), this inconsistency is also apparent: all of these numbers are significantly overpredicted (by factors $\sim$10–100) by the low-κ models in B19, so faster transport is clearly required. In short, the difference between our conclusions (here and in Papers I and II), and those in Pfrommer et al. (2017) and B19, are driven almost entirely by how those authors compare to the observations, rather than by theoretical or numerical differences.

This paper has been typeset from a TEX/LATEX file prepared by the author.

[25] If we use the identical adopted parameters from Pfrommer et al. (2017), their predicted γ ray emission per unit volume in their band 0.1–100 GeV is $\dot{e}_\gamma = 5.67\times10^{-17}\,n_n\,e_{\rm cr}$, so their $L_\gamma^{0.1-100} = \int \dot{e}_\gamma\,{\rm d}^3\boldsymbol{x}$, while $\dot{E}_{\rm cr} = 3.5\times10^{40}\,{\rm erg\,s^{-1}}\,(\dot{M}_*/{\rm M_\odot\,yr^{-1}})$, and therefore in quasi-steady state (when $L_\gamma \ll L_{\rm calor}$), they must have $X_s^\infty \approx 380\,{\rm g\,cm^{-2}}\,(L_\gamma/10^{40}\,{\rm erg\,s^{-1}})\,(\dot{M}_*/{\rm M_\odot\,yr^{-1}})^{-1}$ (for their quoted values of L_γ and $\dot{M}_*$). As $L_\gamma \rightarrow L_{\rm calor}$, of course, $X_s \rightarrow \infty$.